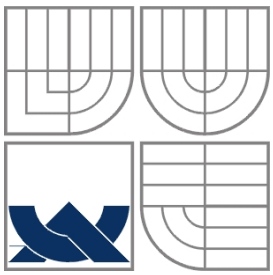
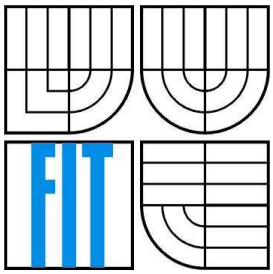


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

VÝVOJ META-SERVERU PRO PŘEDPOVĚĎ VLIVU MUTACÍ NA FUNKCI PROTEINU

DEVELOPMENT OF META-SERVER FOR PREDICTION OF MUTATIONS EFFECTS ON PROTEIN FUNCTION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — Peter Lisák
AUTHOR

VEDOUCÍ PRÁCE — Ing. Petr Jaša
SUPERVISOR

BRNO 2009

Abstrakt

Tato bakalářská práce se zabývá analýzou genomických dat, konkrétně předpovědí vlivu mutací na funkci proteinů z jejich sekvence a terciární struktury. V teoretickém úvodu práce shrnuje základy genetiky a bioinformatiky. Ve výsledkové části se práce zaměřuje na predikční nástroje SIFT, MAPP a AUTO-MUTE. Navrhuje způsob jednotného rozhraní pro práci s těmito nástroji. Společné rozhraní umožňuje zadávat výpočty a sbírat výsledky různých nástrojů z jednoho místa. Ze získaných dat predikuje vlastní konsensuální výsledek s očekávanou vyšší přesností než jednotlivé nástroje. Závěr práce je věnovaný testování aplikace na skutečných datech a porovnání předpovědí s experimentálně získanými výsledky.

Klíčová slova

bioinformatika, genetika, protein, mutace, substituce, predikční nástroj, konsenzus, SIFT, MAPP, AUTO-MUTE, LLWS, meta-server,

Abstract

This bachelor thesis deals with analysis of genomic data, more specifically prediction of effects of mutations on protein function using a protein sequence or tertiary structure. The theoretical introduction describes the basics of genetics and bioinformatics and is followed by description of selected prediction tools such as SIFT, MAPP and AUTO-MUTE. A unified interface for work with different tools is proposed in the thesis. The meta-server interface allows running a computation and collecting results from one site. Meta-server combines results of implemented tools and provides a consensual prediction, which is expected to be more accurate than the results from individual tools. Finally, testing of meta-server on the real data and comparisons of predictions with the experimentally obtained results are presented.

Keywords

bioinformatics, genetics, protein, mutation, substitution, prediction tool, consensus, SIFT, MAPP, AUTO-MUTE, LLWS, meta-server,

Citace

Lisák Peter: Vývoj meta-servera pre predpoveď vplyvu mutácií na funkciu proteínu, bakalářská práce, Brno, FIT VUT v Brně, 2009

Vývoj meta-servera pre predpoveď vplyvu mutácií na funkciu proteínu

Prohlášení

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením Ing. Petra Jaši. Ďalšie informácie mi poskytli Mgr. Antonín Pavelka, Mgr. Eva Chovancová a doc. Jiří Damborský. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

……
Peter Lisák
19. 5. 2009

Poděkování

Týmto smerom by som rád poďakoval všetkým tým, za pomoci ktorých táto práca mohla vzniknúť. Konkrétne pánovi Ing. Petrovi Jašovi za ústretovosť pri konzultáciách, Mgr. Eve Chovancovej a doc. Jiřímu Damborskému za ochotné zasvätenie do problematiky bioinformatiky, Mgr. Antonínovi Pavelkovi za pomoc z technickými a programátorskými problémami a Bc. Rostislavovi Wolnému za spoluprácu. Samozrejme všetkým ostatným za morálnu podporu.

Obsah

1 Úvod

Zadanie tejto práce vzniklo v spolupráci s bioinformatickým tímom vedením Mgr. Evou Chovancovou. Tento tím je jedným zo štyroch vo výskumnej skupiny proteínového inžinierstva (angl. Protein Engineering Group) Loschmidtových laboratórií Masarykovej univerzity v Brne vedenej doc. Jiřím Damborským. Vzhľadom na rozsah zadania a množstvo problematiky a nástrojov k naštudovaniu bolo už zo zadania doporučené ho riešiť vo dvojici. Takže téma bola vypracovaná v spolupráci s Rostislavom Wolným študentom piateho ročníka Fakulty informatiky Masarykovej univerzity.

Cieľom tejto bakalárskej práce je sprehľadniť a uľahčiť prácu genetikov a proteínových inžinierov snažiacich sa predpovedať vplyv mutácie v DNA na funkciu proteínu. Za poslednú dekádu vzniklo za týmto účelom množstvo predikčných nástrojov s viac či menej rozdielnym prístupom k predikcii. V tejto práci preto usilujeme o vytvorenie pre užívateľa centralizovaného nástroja – meta-servera, ktorý ponúkne jednotné rozhranie a istý štandard pre prácu s týmito nástrojmi a z ich výsledkov vytvorí konsenzus, čím *de facto* vznikne nový predikčný nástroj. Konsenzuálny výsledok tvorený z výsledkov nástrojov s rozdielnym prístupom k predpovedí robí meta-server presnejším a dôveryhodnejším zdrojom predikcií.

Text k tejto práci sa skladá zo šiestich kapitol a postupne čitateľov zoznámi s riešenou problematikou od biologických základov až po samotnú implementáciu meta-servera. Teoretická časť práce má za úlohu pomôcť čitateľovi preniknúť do problematiky. Poskytnúť mu rýchlo kurz genetiky, oboznámiť ho s základnými makromolekulovými látkami v organizmoch - molekulou DNA a bielkovinami. Dôležité je pochopiť význam a dopad jednonukleotidových mutácií na funkciu proteínov a následne ich dopad na samotnú kvalitu života organizmu. Ďalej si objasníme prínos informatiky a výpočtovej techniky v genetike. Zoznámime sa s možnosťami udržiavania dát v súčasných nadnárodných biologických databázach a ich základným spracúvaní ako vytváranie zarovnania proteínových sekvencií alebo evolučných stromov.

V kapitole príznačne nazvanej Nástroje a možnosti predikcie sa zoznámime s možnosťami predikcie vplyvu mutácii *in silico*, tj. na počítači. Popíšeme si nástroje SIFT, MAPP a AUTO-MUTE, ktoré sú v súčasnej dobe použité v meta-servery. Ukážeme si prácu s týmito nástrojmi, zanalyzujeme ich komunikačné rozhranie, stanovíme požadované vstupy metód a rozoberieme štruktúru ich výstupov.

V nasledujúcej kapitole sa dostaneme k vývoju samotného meta-servera. Na začiatku je návrh aplikácie, venovaný hlavne spracovaniu vstupných hodnôt a začiatku výpočtu. V ďalšej časti sa potom kapitola venuje implementácii návrhu. V podkapitolách stručne prezentuje výsledky testovania a konkretizuje nedostatky aplikácie. Popisuje testovanie meta-servera a porovná výsledky

meta-servera so skutočnými, ale i predpovedanými dátami iných nástrojov. Na základe testovania vyhodnotíme relevantnosť nami navrhnutej konsenzuálnej predpovede.

Úplným záverom sa navrhnú úpravy a vylepšenia aplikácie do budúcnosti. Záver taktiež v krátkosti zhrnie prínos tejto práce pre mňa, nové vedomosti, obzory a skúsenosti so spoluprácou.

2 Teória

Téma tejto práce zasahuje aj do iných vedných oblastí ako je samotná informatika. Preto sa v tejto časti budeme zaoberať niektorými disciplínami biológie ako napr. genetikou a hlavne si vysvetlíme problematiku vzniku mutácii v DNA a ich vplyvu na vznik proteínov a samotnú funkčnosť proteínov. Ďalej si povieme niečo o využití počítačovej technológie v tejto oblasti.

2.1 Genetika

Biológia, ako veda študujúca organizmy a všetko čo s nimi súvisí, zastrešuje viaceré vedné disciplíny. Jednou z nich je aj genetika, zaoberajúca sa dedičnosťou a premenlivosťou živých organizmov. Názov pochádza z gréckeho *genno* (plodím, rodím), ale tiež súvisí so slovom gén, čo označuje jednotku dedičnej informácie [1].

Patrí k mladším biologickým disciplínam. Jej počiatky siahajú do 19. storočia. Svoj rozmach zažíva od 2. polovice 20. storočia až do dnes a s využitím počítačov sa dá predpokladať aj jeho pokračovanie v ďalšej budúcnosti. Veľký prínos pre človeka má hlavne lekárska alebo klinická genetika, ale tak tiež genetika na poli šľachtenia zdokonaľovania rastlinných a živočíšnych druhov.

2.1.1 Historický vývoj genetiky

Základy genetiky položil Gregor Johann Mendel (1822-1884) v druhej polovici 19. storočia. Ako augustiánsky mních kláštora v Brne skúmal dedičné znaky pri krížení hybridov hrachu (*Pisum sativum*). Pozoroval 7 dedičných znakov (tvar semien a luskov, zafarbenie delôh, kvetov a nezrelých luskov, dĺžku stonky a postavenie kvetov), z ktorých pomocou matematického vyhodnotenia výsledkov sformuloval základné zákony genetiky. V roku 1866 vydáva prácu o svojom pozorovaní s názvom Versuche über Pflanzenhybriden (Pokusy s rastlinnými krížencami). Avšak ako mnoho prác iných vedcov ani jeho práca nemala ohlas a bola jeho súčasníkmi uvrhnutá do zabudnutia. Začiatkom 20. storočia bola Mendelova práca znovu objavená a jeho objavy potvrdené. Dochádza k vzniku genetiky ako plnohodnotného vedného oboru. Samotný názov genetika roku 1906 zaviedol anglický profesor William Bateson (1861 – 1926), ďalej i pojmy heterozygót a homozygót. S dánskym vedcom Wilhelm Johannsen (1857 – 1927) prichádzajú pojmy ako gén, genotyp a fenotyp. V roku 1933 sa prvým genetikom, ktorý za svoje objavy získal Nobelovu cenu, stáva Američan Thomas Hunt Morgan (1866 - 1945). Venoval sa štúdiu chromozómov na modelovom organizme – octomilke (*Drosophila melanogaster*), priniesol tým množstvo nových poznatkov o génoch a génovej väzbe. Objav DNA a jej úlohy ako nositeľky genetickej informácie dala podnet k ďalšiemu rozvoju vedy. Prelomovým sa stáva rok 1953, kedy sa americkým vedcom Jamesovi D.

Watsonovi a Francisovi H. Crickovi podarilo popísať a vytvoriť prvý štruktúrny model dvojzávitnice DNA. Za svoj prínos v oblasti genetiky získavajú v roku 1962 Nobelovu cenu spolu s Mauricom H. F. Wilkinsonom, ktorý prispel röntgenovou štúdiou DNA [2].

Týmito objavmi bol spustený obrovský rozmach genetiky, ktorý napredoval cez potvrdenie tripletového kódu, stanovenie počtu chromozómov a sekvenovanie genómov jednoduchých organizmov. Doterajším vyvrcholením je sekvenovanie ľudského genómu (2003 – kompletná sekvencia). V súčasnej dobe prebiehajú výskumy zamerané na ľudský génom, napr. v oblasti farmakológie alebo génovej terapie, cieľom je personifikovaná medicína. Využitím bioinformatických technológii a praktík sa ešte viac urýchľuje intenzita výskumu.

2.1.2 Molekulárna genetika

Molekulárna genetika je oblasť genetiky, ktorá sa sústredí na štúdium štruktúry, funkcie a variability génov na molekulárnej úrovni. Je vývojovým pokračovaním klasickej genetiky. Základne tézy boli získane experimentmi na mikroorganizmoch a neskôr potvrdené na vyšších živočíchoch [1].

V bunkách rastlín a živočíchov sa nachádza množstvo makromolekulových látok nevyhnutných pre jej fungovanie. Jedná sa hlavne o nukleové kyseliny alebo proteíny. V ďalších podkapitolách si obe bližšie makromolekuly priblížime.

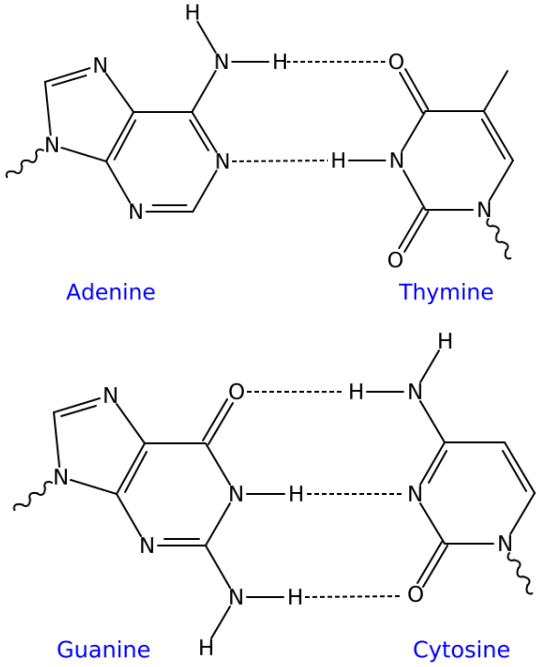
2.1.3 DNA

Deoxyribonukleová kyselina, skratka DNK alebo DNA (angl. deoxyribonucleic acid), je u väčšiny organizmov nositeľkou genetickej informácie. Jej úlohou je uchovať a odovzdať genetickú informáciu (určuje poradie aminokyselín v proteínoch). V globále určuje celý vývoj a vlastnosti organizmu. DNA patrí medzi nukleové kyseliny. Nukleové kyseliny sú tvorené polynukleotidovými reťazcami, ktoré sa skladajú z nukleotidov [1].

V eukaryotických bunkách, čiže v bunkách rastlinných alebo živočíšnych, je DNA uložená v chromozómoch vo vnútri jadra bunky a má tvar špirály – dvojzávitnice. Taktiež sa nachádza v dvoch organelách týchto buniek a to v mitochondriách a chloroplastoch. Bezjadrové organizmy, tzv. prokaryoty, teda hlavne baktérie, majú DNA uloženú voľne v cytoplazme, kde ju označujeme ako nukleotid alebo kruhový chromozóm. Tieto bunky môžu obsahovať taktiež doplnkovú genetickú výbavu v tzv. plazmidoch. Plazmidy nie sú pre prokaryoticku bunku životné dôležité, ale dajú sa využiť v génovom inžinierstve na vnášanie cudzorodej genetickej informácie do buniek geneticky nepríbuzných organizmov, napr. plazmid baktérie obsahujúci gén na tvorbu ľudského inzulínu umožnil jeho veľkoobjemovú produkciu a tak pomohol diabetikom plnohodnotnejšie žiť. Kompletná genetická informácia daného organizmu sa nazýva génom [2].

2.1.3.1 Nukleotid

Základnou stavebnou jednotkou nukleových kyselín sú nukleotidy. Ide o molekulu zloženú z dusíkatej heterocyklickej zlúčeniny (tzv. báza), päťuhlíkatého cukru (2-deoxyriboza u DNA, ribóza u RNA) a fosfátovej skupiny. DNA obsahuje 4 typy bázy (Obrázok 1), 2 pyrimidíny - cytozín (C) a tymín (T) a 2 puríny – guanín (G) a adenín (A). V RNA je tymín nahradený uracilom. Nukleotidová báza dáva meno celému nukleotidu. Jednotlivé nukleotidy sú spojené kovalentnou fosfodiestérovou väzbou v polynukleotidový reťazec. Lineárne poradie nukleotidov v reťazci kóduje dedičnú informáciu. Genetická informácia je uložená tzv. tripletovým kódom. Jedno kódové slovo – kodón - je určený tripletom (3 po sebe idúce páry nukleotidov) [1; 2; 3].



Obrázok 1 Pyrimidinové bázy cytozín a thymín a purinové adenín a guanín a ich párovanie [1]

2.1.3.2 Štruktúra

DNA je vytvorená z dvoch polynukleotidových reťazcov vzájomné prepletených do dvojzávitnice. Obe vlákna dvojzávitnice sú pospájané pomocou vodíkových väzieb, pričom platí princíp tzv. komplementarity báz. Komplementarita je jedna z dôležitých vlastností DNA a znamená, že sa párujú len dve špecifické dusíkaté bázy. Všeobecne teda purínové s pyrimidinovými, guanín s cytozínom pomocou troch vodíkových väzieb a adenín s tymínom pomocou dvoch vodíkových väzieb (Obrázok 1)[1].

Primárna štruktúra je daná poradím nukleotidov (sekvenciou) polynukleotidového reťazca a je ním určená genetická informácia. Nukleotidy sú pospájané väzbou medzi fosfátovým zvyškom naviazaným na 5’ uhlíkovom atóme deoxyribozy jedného nukleotidu a hydroxilovou skupinou druhého nukleotidu viazanou na 3’ uhlíkovom atóme deoxyribozy. Takto vzniká fosfodiesterová

väzba. Rozlišujeme dva konce polynukleotidového reťazca: koniec 5’ ukončený fosfátom a koniec 3’ ukončený hydroxilovou -OH skupinou.

Sekundárna štruktúra označuje priestorové usporiadanie dvoch DNA reťazcov. Reťazce sú komplementárne pospájané, vzájomne sú pospájané v opačnom poradí. Jeden je orientovaný v smere 5’ →3’ a druhý 3’ →5’. Toto dvojvlákno sa stáča do pravotočivej závitnice – α-helixu. Forma stočenia nie je vždy a za každých podmienok rovnaká:

i. Ds forma A-pravotočivá
ii. Ds forma B-pravotočivá
iii. Ds forma Z-ľavotočivá

Terciárnu štruktúru určuje priestorové usporiadanie dvojzávitnice. Tá sa stočí do superhelixu. Takto zvinutá DNA sa nazýva superšpiralizovaná DNA.

2.1.3.3 Gén

Gén je považovaný za jednotku genetickej informácie. Gén je úsek molekuly DNA, ktorý má špecifickú funkciu. Podľa funkcie môžeme gény rozdeliť na: štruktúrne (i) – kódujúce štruktúru proteínu, regulačné (ii) – oblasti DNA rozpoznávané špecifickými proteínmi, ktoré regulujú expresiu génu a RNA gény (iii) – syntetizujú sa podľa nich molekuly transférových RNA (tRNA) a ribozomálnych RNA (rRNA). Podľa účinnosti ich môžeme rozdeliť na monogénné gény, ktoré majú veľký účinok a na samotnej tvorbe určitého dedičného znaku sa podieľa málo génov, väčšinou len jeden. V tomto prípade sa jedná o znak kvalitatívny. A polygénné gény, ktoré majú malý účinok a na tvorbe znaku sa podieľajú vo väčšom počte a ovplyvňujú kvantitatívne znaky [3].

2.1.3.4 Replikácia

Aby si DNA udržala funkciu prenosu genetickej informácie, musí byť schopná zdvojenia – replikácie. Táto schopnosť je nenahraditeľná pri delení buniek, aby dcérska bunka dostala plnohodnotnú genetickú informáciu. Pri zdvojovaní vznikajú 2 identické dcérske DNA reťazce z jedného materského. Každý z týchto reťazcov obsahuje jedno vlákno pôvodnej DNA – tzv. semikonzervatívny postup. Jedná sa o enzymaticky riadený proces kopírovania sekvencie DNA na základe komplementarity. Enzymaticky riadený znamená, že dôležitú úlohu v tomto procese zohrávajú enzýmy (napr. DNA polymerázy). Tieto enzýmy vždy replikujú DNA od určitého miesta. Prokaryoty majú na molekule jeden replikačný bod, eukaryoty ich majú viacero, aby replikácia väčších genómov netrvala dlho. Vlákna DNA sa najskôr začnú rozpletať a vzďaľovať od seba. Na oddelené vlákna sa komplementárne začnú prikladať a naväzovať nukleotidy, postupuje sa smerom 5’ →3’. Preto sa vždy jedno vlákno dosyntetizuje neskôr z dôvodu, že sa postupuje v protismere rozpletania. Možnosť chyby je jedna na 10^7 zreplikovaných bázy (teda teoreticky môžu vzniknúť aj iné dvojice ako A-T a G-C, ale sú menej stabilné), niektoré DNA-polymerázy majú navyše korekčnú

funkciu. Takže k chybe behom replikácie dochádza len vzácne, ale i tak môže dôjsť k trvalému zafixovaniu zmeny v DNA. Táto zmena sa nazýva mutácia [3].

2.1.3.5 Transkripcia a translácia

Prenos dedičnej informácie z DNA do štruktúry proteínov (expresia génov) prebieha v dvoch stupňoch, ktorými sú transkripcia a translácia [4].

Transkripcia je prvý stupeň expresie génu. Behom transkripcie dochádza k prepisu genetickej informácie z DNA do mRNA. Sprostredkovaná je enzýmom RNA polymerázou, ktorá syntetizuje jeden komplementárny reťazec primárneho transkriptu z DNA templátu. Prepis je iniciovaný, potom ako sa RNA polymeráza naviaže na promotorovú sekvenciu DNA (špecializovaná DNA sekvencia blízko začiatku génu, kde začína transkripcia) a rozvine jej dvojzávitnicu, aby sa mohli vystaviť báze templátu k bázam tvoreného reťazca RNA.

Translácia je preklad genetickej informácie z poradia nukleotidov na mRNA do poradia aminokyselín v polypeptidovom reťazci. Behom translácie je sekvencia nukleotidov mRNA postupne čítaná po trojiciach nukleotidov – kodónoch (Obrázok 2). Slúžia k tomu molekuly tRNA, krátke jednoreťazcové RNA molekuly, ktoré na jednom konci nesú špecifickú trojicu nukleotidov – antikodón, na druhom konci má potom naviazanú odpovedajúcu aminokyselinu (pomocou enzýmu aminoacyl-tRNA-synthetázy). Kodón mRNA je rozpoznaný molekulou tRNA s komplementárnym antikodónom. Dochádza k párovaniu daného úseku mRNA a tRNA a aminokyselina nesená danou tRNA je začlenená na príslušné miesto vznikajúceho polypeptidového reťazca. Každý kodón kóduje jednu aminokyselinu. Existuje 64 trojprvkových kombinácii a keďže aminokyselín je len 20, niektoré z nich sú kódované niekoľkými rôznymi kodónmi. K translácii dochádza na ribozómoch (veľké komplexy rRNA a proteínov).

	U		C		A		G	
U	UUU	fenylalanin	UCU	serin	UAU	tyrosin	UGU	cystein
	UUC	fenylalanin	UCC	serin	UAC	tyrosin	UGC	cystein
	UUA	leucin	UCA	serin	UAA	**stop**	UGA	**stop**
	UUG	leucin	UCG	serin	UAG	**stop**	UGG	tryptofan
C	CUU	leucin	CCU	prolin	CAU	histidin	CGU	arginin
	CUC	leucin	CCC	prolin	CAC	histidin	CGC	arginin
	CUA	leucin	CCA	prolin	CAA	glutamin	CGA	arginin
	CUG	leucin	CCG	prolin	CAG	glutamin	CGG	arginin
A	AUU	izoleucin	ACU	treonin	AAU	asparagin	AGU	serin
	AUC	izoleucin	ACC	treonin	AAC	asparagin	AGC	serin
	AUA	izoleucin	ACA	treonin	AAA	lysin	AGA	arginin
	AUG	**metionin**	ACG	treonin	AAG	lysin	AGG	arginin
G	GUU	valin	GCU	alanin	GAU	kys.	GGU	glycin
	GUC	valin	GCC	alanin	GAC	asparagová	GGC	glycin
	GUA	valin	GCA	alanin	GAA	kys.	GGA	glycin
	GUG	valin	GCG	alanin	GAG	glutamová	GGG	glycin

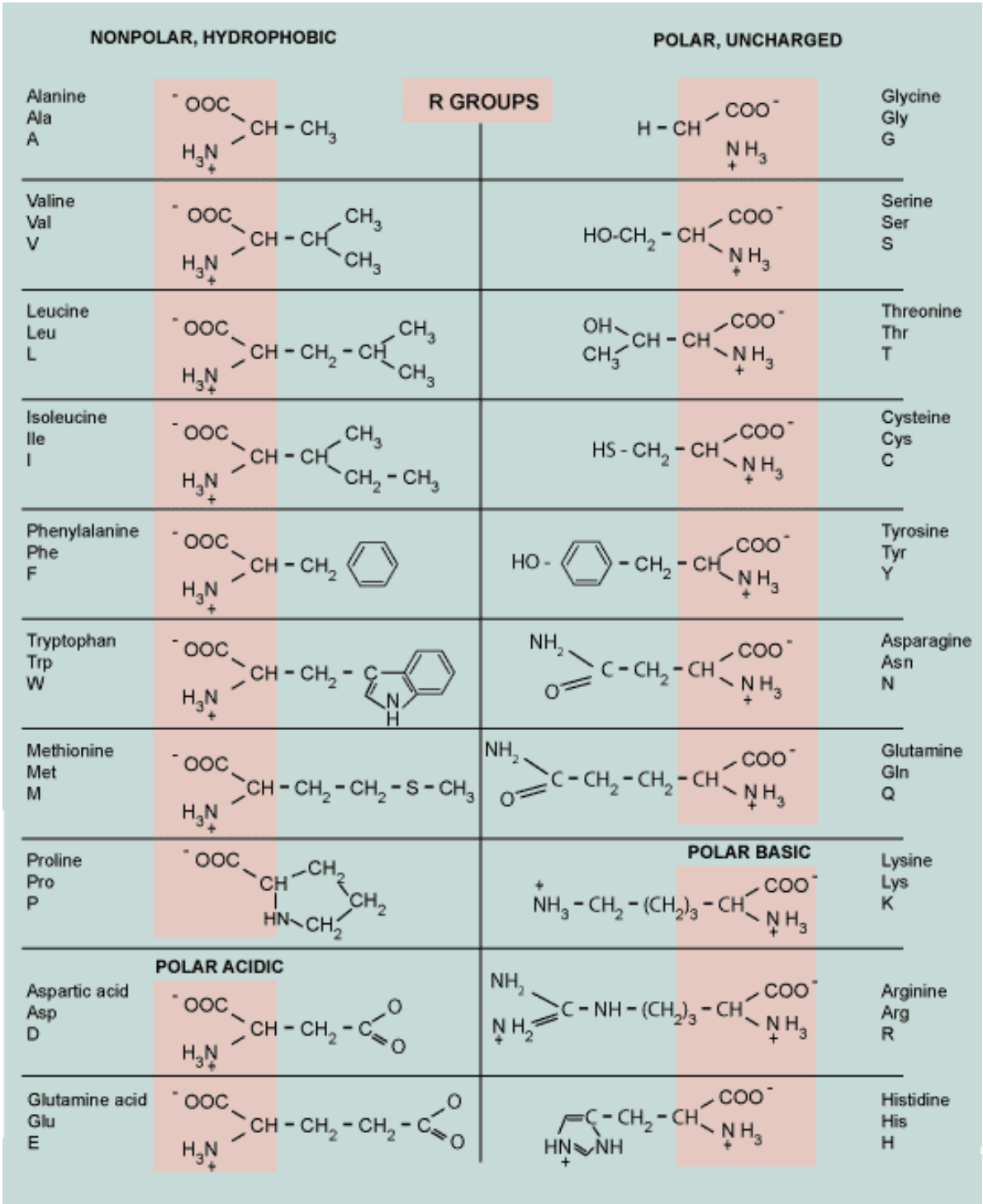
Obrázok 2 Tripletový genetický kód [2]

2.1.4 Proteíny

Proteíny alebo bielkoviny sú organické zlúčeniny zložené z aminokyselín spojených tzv. peptidickými väzbami v polypetidové reťazce. Názov pochádza z gréckeho proteios a znamená základný, prvotný. Sekvencia aminokyselín v polypeptidovom reťazci je dekódovaná z genetickej informácie DNA. Proteíny sú jednou zo životne dôležitých zlúčenín v živých organizmoch. Vykonávajú rôzne úlohy, ako transportnú (hemoglobín), stavebnú (kolagén), katalytickú (enzýmy), regulačnú (transkripčné faktory) alebo imunitnú (imunoglobulín) [1; 3].

2.1.4.1 Aminokyseliny

Z hľadiska molekulárnej biológie sa aminokyselinou rozumie 20 základných stavebných zložiek všetkých bielkovín (Obrázok 3). Sú označované taktiež ako proteínogénne aminokyseliny. V živých organizmoch sú súčasťou bielkovín, peptidov a mnohých hormónov. Rastliny dokážu syntetizovať všetky aminokyseliny z anorganických látok, živočíchy si dokážu tvoriť len niektoré aminokyseliny, iné (esenciálne aminokyseliny) musia získavať z potravy [1].



Obrázok 3 Proteínogénne aminokyseliny[5]

2.1.4.2 Štruktúra

Štruktúra proteínov sa podľa úrovne pohľadu delí na primárnu, sekundárnu, terciárnu a u niektorých kvartérnu. Obecne sa bielkovina skladá do tvaru s najnižšou voľnou energiou [1].

Primárna štruktúra je daná sekvenciou aminokyselín v polypeptidovom reťazci. Tá je, ako už bolo spomínane predtým, uložená v DNA tripletovým kódom. Sekvencia určuje vlastnosti a štruktúru bielkoviny.

Sekundárna štruktúra, terciárna a kvartérna je daná priestorovým usporiadaním polypeptidového reťazca. U sekundárnej sa jedná o usporiadanie na „krátke vzdialenosti“, teda medzi niekoľkými po sebe idúcimi aminokyselinami – α-závitnice (pravotočivá závitnica) a β-štruktúra (skladaný list). Keďže toto usporiadanie je lokálne, v jednej molekule sa môže vyskytnúť viacero rôznych sekundárnych štruktúr.

Terciárna štruktúra udáva priestorové usporiadanie celého reťazca – celkový tvar molekuly. Určuje základné funkcie bielkoviny.

Kvartérna štruktúra je daná vzájomným pôsobením viacerých proteínových štruktúr. Definuje sa u tzv. proteínových komplexoch, ktoré sú zložené z viacerých molekúl.

2.1.4.3 Enzýmy

Bielkoviny, ktorých funkciou je katalyzovať - urýchľovať chemickú reakciu znižovaním jej aktivačnej energie. Majú kľúčovú úlohu v metabolizme živých organizmov, katalyzujú biochemické reakcie od trávenia až po kopírovanie DNA[2]. Produkt jedného enzýmu sa môže stať substrátom ďalšieho enzýmu. Jednotlivé enzýmy sú určené pre isté druhy reakcie. Ich účinky dokážu v rádoch miliónov urýchliť reakcie, napr. u dekarboxylázach to je až 10^{17} násobne urýchlenie, tj. katalyzovaná reakcia potrvá len 18ms, ale bez enzýmu by to bolo 78 miliónov rokov [1].

Napriek tomu, že sú enzýmy typicky tvorené 100-300 aminokyselinami, len malá časť z týchto aminokyselín (3-4 v priemere) sa priamo účastní katalýzy – označujeme ich ako katalytické aminokyseliny. Nachádzajú sa v aktívnom mieste enzýmu, čo je to miesto, kde sa viaže substrát a dochádza ku katalýze chemickej reakcie [1].

2.1.5 Mutácie

Po uchovaní a prenose genetickej informácie je jej zmena ďalšou podmienkou k evolúcii. V genetike sa mutáciami rozumejú zmeny v genetickej informácii v DNA. Vznikajú z veľkej väčšiny náhodne a väčšinou vplyvom vonkajších mutagénnych faktorov – indukované mutácie. K spontánnym mutáciám dochádza zriedka kedy napr. pri replikácii DNA [3].

2.1.5.1 Typy mutácii

Mutácie je možné rozdeliť podľa viacerých kritérií. Z hľadiska lokalizácie na génové (zmeny v sekvencii DNA), chromozomálne (zmena štruktúry a tvaru chromozómu), genómové (zmena počtu chromozómov) a nechromozomálne (mimojadrové). Podľa typu sekvencie v molekule DNA rozdeľujeme na mutácie v kódujúcich sekvenciách a nekódujúcich [2].

Mutácie v kódujúcej časti sekvencie DNA často ovplyvňujú kvalitu samotného proteínu zakódovaného v tejto časti sekvencie. Mutácie v nekódujúcej časti môžu ovplyvniť reguláciu a kvantitu zakódovaného proteínu. Kódujúca časť génu tvorí menej ako 5% sekvencie, takže k ním dochádza menej, ale za to sa u nich dá očakávať väčší vplyv na biologickú funkciu proteínu. Mutácie

v kódujúcej časti sekvencie sa ďalej delia na synonymné a nesynonymné mutácie. Pri synonymných k žiadnej zámene zakódovanej aminokyseliny nedôjde. Dôvodom je, že tripletový kód má 64 kombinácii a aminokyselín je len 20. Nesynonymné mutácie, sú mutácie v kódovej časti génu, ktoré môžu buď meniť zmysel alebo byť tzv. nezmyselná, ktorá predčasne ukončí preklad. Každopádne ovplyvňujú primárnu štruktúru, ale nie je isté ako ovplyvnia priestorové usporiadanie proteínu a ako jeho funkciu a aký to vo výsledku bude mať vplyv na samotný organizmus [1; 3].

Mutácia génu je dedičná zmena v sekvencii nukleotidov v DNA. Majú niekoľko mechanizmov vzniku a väčšinou vznikajú pri replikácii DNA [2].

i. delecia – strata jedného alebo viacerých párov nukleotidov, skracuje sekvenciu
ii. inzercia – včlenenie jedného alebo viacerých párov nukleotidov do sekvencie DNA, predlžuje sekvenciu
iii. inverzia – prevrátenie poradia dvoch alebo viacerých po sebe nasledujúcich párov nukleotidov
iv. substitúcia – nahradenie jedného alebo niekoľkých párov nukleotidov inými:
tranzícia – purínový nukleotid za purínový (A ←→ G), pyrimidínový za pyrimidínový (T ←→ C) a tranverzia – purínový nukleotid za pyrimidínový (A, G ←→ T, C)

2.1.5.2 Prirodzené mutácie

Genetické choroby a nádorové bujenie sú spôsobované práve mutáciami. Touto oblasťou sa zaoberá lekárska alebo klinická genetika. Uplatňuje sa pri skúmaní rôznych genetických chorôb, ich početnosti a genetickej determinácii istých ľudských znakov. Tiež sa už stále častejšie môžeme stretnúť s génovou terapiou a genetickým poradenstvom, najmä pri plánovaní potomkov a prevencii vrodených vývojových vád. Ďalej tiež zohráva významnú rolu vo foréznych metódach v kriminalistike, hlavne DNA testy pri usvedčovaní zločincov, identifikácii telesných pozostatkov alebo stratených osôb [1].

Jednonukleotidový polymorfizmus- SNP

SNP (angl. single nucleotide polymorphism, čítaj snip) sú jednonukleotidové polymorfizmy sekvencie DNA. Teda variácia DNA, ktorá v sekvencii vzniká zámenou jedného nukleotidu [2].

Množstvo SNP nemá žiadny efekt na funkciu organizmu, iné sú spájané s predizpozíciami k chorobám či odpoveďou organizmu na liečbu a iné environmentálne faktory. Teda je možné vyžitie SNP profilov v biomedicínskom výskume, vývoji farmaceutických produktov či lekárskej diagnostike. SNP sú považované za kľúč k personifikovanej medicíne, kde podľa rozboru pacientových génov a SNP profilu sa pacientovi aplikuje liečba „na mieru“ [6]. SNP sa v géne môže vyskytovať v kódovej alebo nekódovej časti sekvencie. V rámci tejto práce nás budú zaujímať

jednonukleotidové mutácie v kódovej časti, ktoré zmenia výsledný polypeptidový reťazec - nesynonymné SNP (nsSNP).

nsSNP môžu ovplyvniť funkciu proteínu, čím majú najväčší dopad na ľudské zdravie v porovnaní s SNP v iných miestach genómu. Preto je dôležité roztriediť tie nsSNP, ktoré ovplyvnia funkciu proteínu od tých, ktoré sú funkčne neutrálne. V ľudskej populácii je odhadované, že má 67 000 – 200 000 spoločných nsSNP a každá osoba ich má asi 24 000 – 40 000. Bolo by časovo a finančne veľmi náročné experimentálne odhaliť vplyv každej nsSNP na funkciu proteínov. Bioinformatické výpočtové metódy môžu napomôcť vybrať z tohto množstva funkčne významné aminokyselinové substitúcie a prioritizovať ich ďalšie štúdium [7].

2.1.5.3 Proteínové inžinierstvo

Rôznymi kombináciami aminokyselín v sekvencii bielkovín môže vzniknúť ďaleko viac proteínových štruktúr ako sa nachádza v živých organizmoch na Zemi. Navrhovaním a konštrukciou týchto nových, pozmenených proteínov sa zaoberá proteínové inžinierstvo [8].

Kritickým krokom proteínového inžinierstva je navrhnúť také mutácie, ktoré povedú ku chceným zmenám vo vlastnostiach študovaného proteínu. Predpokladom pre racionálny design mutácií je dobrá znalosť vzťahov medzi štruktúrou a funkciou daného proteínu [9]. Alternatívnym prístupom proteínového inžinierstva sú techniky riadenej evolúcie, kde sú mutácie do proteínu vnášané náhodne a pripomínajú urýchlenú evolúciu [10]. Oba prístupy, racionálny design proteínov a riadenú evolúciu, je možné taktiež kombinovať [11].

Keďže experimentálna práca v laboratóriu je časovo i finančne veľmi náročná, a preto je dobré čo najviac minimalizovať možnosť neúspechu experimentu, tj. možnosť, že mutácia nebude mať očakávaný vplyv na funkciu. Pomocou vhodných výpočtových metód je možné najskôr predikovať vplyv mutácii na stroji, kde je to finančne a časovo nenáročne. Experimentálne sa potom skontruštruujú len tie proteíny, ktoré pri simulácii na počítači mali očakávané výsledky.

2.2 Bioinformatika

Bioinformatika je hraničná disciplína viacerých vedných odborov: informatiky, biológie, biochémie a aplikovanej matematiky. Rieši biologické problémy zvyčajne na molekulovej úrovni. Jej konečným cieľom je odhaliť podstatu biologických informácii ukrytých v množstve dát, čím sa ozrejmia základné biologické funkcie organizmov. Bioinformatiku môžeme chápať ako vedu zaoberajúcu sa získavaním, spracovaním a analýzou dát o sekvencii a štruktúre biologických makromolekúl. Taktiež ju môžeme chápať i vo význame širšom ako „využitie počítačov k hľadaniu odpovedí na biologické otázky“ [12].

Bioinformatika pracuje *in silico*, tj. pracuje či simuluje na počítači, na rozdiel od práce na živých organizmoch, či v ich častiach v prirodzených (*in vivo*) alebo laboratórnych (*in vitro*) podmienkach. Počítače slúžia pre hromadenie, ukladanie, analýzu a prepojenie biologických dát.

2.2.1 Bioinformatické databáze

Bioinformatické databáze pracujú hlavne s dátami ako sekvencie proteínov a nukleových kyselín, štruktúra makromolekúl (trojrozmerný model hlavne proteínov), údaje o aktivite a expresii génov, údaje o funkciách génov a ich produktov, údaje o interakciách medzi proteínmi a DNA [12].

Sekvenčné databáze

Sekvenčné informácie z molekuly DNA alebo proteínu sú vo svojej podstate digitálne – tieto molekuly môžeme chápať aj ako digitálne médium. Postupnosť nukleotidov v DNA prevádzame do digitálnej formy pomocou tzv. sekvenovania DNA. Sekvencie proteínov sú vo väčšine prípadov odvodené teoreticky zo sekvencie príslušných DNA. Sekvenčné dáta digitálne zapisujeme do postupnosti jednoznačných znakov odpovedajúcich jednotlivým typom monomérov (nukleotidov alebo aminokyselín) po rade tak ako sú v molekule pri postupe v smere odpovedajúcom smeru biosyntézy daného typu molekuly. Pre DNA to je v smere 5’ → 3’ a u proteínov od N-konca k C-koncu. K označeniu monomérov sa väčšinou používajú jedno písmenkové kódovanie monomérov stanovené Medzinárodnou úniou pre čistú a aplikovanú chémiu (IUPAC) [12].

Pre zápis a identifikáciu sekvencií sa používa viacero formátov. Základným je surová forma – (angl. raw data) – reťazec znakov. Nevýhodou je málo prídavných informácii o proteíne. Jedným z najpoužívanejších formátov je FASTA. Prvý riadok „hlavička“ tohto formátu začína znakom „>“ a obsahuje názov sekvencie, anotáciu a iné. Ďalšie riadky sú samotná sekvencia (60 znakov na riadok) [12].

```
>gi|2506562|sp|P03023|LACI_ECOLI   LACTOSE OPERON REPRESSOR
MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQ
SLLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVS
GLIINYPLDDQDAIAVEAACTNVPALFLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQ
QIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREGDWSAMSGFQQTMQMLNEGIVPT
AMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDSSCYIPPLTTIKQDFRLLGQTS
VDRLLQLSQGQAVKGNQLLPVSLVKRKTTLAPNTQTASPRALADSLMQLARQVSRLESGQ
```

V súčasnosti vo svete existujú tri hlavné databáze DNA sekvencii. Európska EMBL (angl. European Moleculary Biology Laboratory Data Library [13]), americká GenBank spravovaná inštitúciou NCBI (National Center for Biotechnology Information [14]) a japonská DDBJ (DNA Data Bank of Japan [15]) tvoria medzinárodné konzorcium a sú vzájomne previazané [16]. Proteínová databáza NCBI [17; 18] a UniProt [19; 20] predstavujú hlavné zdroje proteínových sekvencii. Taktiež existujú databáze, ktoré obsahujú priamo záznamy sekvencii s SNP, napr. SNP databáza spravovaná NCBI [21].

Štruktúrne databáze

Informácie o priestorovej štruktúre proteínov sa získavajú ťažšie ako ich sekvencia, preto i záznamov v štruktúrnych databázach je menej. Priestorovú štruktúru molekuly proteínu je možné určiť napr. pomocou [1]:

- NMR spektroskopie, ktorá využitím interakcie atómových jadier a magnetického poľa rozdelenie energii v jadrách
- röntgenovej štrukturálnej analýzy, ktorá pomocou kryštalických vzorkou určuje 3D model molekuly
- elektrónovej mikroskopie

Informácie o priestorovej štruktúre sa uchovávajú napr. v PDB súboroch alebo PDBML. PDB súbory sú textové súbory s 80 znakmi na riadok (z historických dôvodov), každý riadok je samo identifikovateľný – teda prvých 6 znakov nesie názov záznamu, ktorý môže byť na viacerých riadkoch. PDBML je založený na syntaxi XML súborov. Každá publikovaná štruktúra ma pridelený 4 znakový kód PDB ID – identifikátor konkrétnej štruktúry [22].

```
HEADER    ASPARTYL PROTEASE                          11-MAY-97   1AJV
TITLE     HIV-1 PROTEASE IN COMPLEX WITH THE CYCLIC SULFAMIDE
TITLE    2 INHIBITOR AHA006
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: HIV-1 PROTEASE;
COMPND   3 CHAIN: A, B;
...
```

Proteínovú databanku (angl. Protein Data Bank - PDB) spravuje konzorcium wwPDB (the Worldwide Protein Data Bank), ktorej zakladajúcimi členmi sú RCSB PDB (USA), PDBe (Európa) and PDBj (Japonsko). Jedná sa o centrálne úložisko experimentálne získaných štruktúrnych dát proteínov a nukleových kyselín [23; 24].

2.2.2 Bioinformatické algoritmy

Proteínové databáze obsahujú obrovské množstvo sekvenčných dát, a preto potrebujeme určité mechanizmy, ktoré nám umožňujú rýchle vyhľadávanie v týchto databázach.

Proteínové sekvencie odvodené od rovnakej pôvodnej sekvencie sa nazývajú homológne. Oproti tomu stoja analógne sekvencie – proteíny s podobnými vlastnosťami, ale iného pôvodu. Homológne sekvencie majú dva typy orthology (gény odvodené od rovnakej sekvencie v odlišných živočíšnych druhoch, často plnia rovnakú funkciu) a paralogy (homológne sekvencie vzniklé pôvodne v rámci jedného živočíšneho druhu duplikáciou pôvodnej sekvencie, postupom evolúcie často začínajú plniť odlišnú funkciu). Zarovnanie sekvencii je spôsob, ako identifikovať odpovedajúce si miesta v homológnych sekvenciách. Z digitálneho hľadiska sú zarovnané sekvencie prezentované ako riadky v matici s vloženými medzerami medzi rezíduá, tak aby identické alebo podobné znaky boli

zarovnané v rovnakých stĺpcoch. Využívajú sa dva výpočtové prístupy: globálne zarovnanie a lokálne zarovnanie. Globálne zarovnanie má snahu obsiahnuť každé rezíduum v každej sekvencii. Je v hodné pre veľmi podobné sekvencie so skoro totožnou dĺžkou. Zatiaľ, čo lokálne vyberá podobne miesta na dlhých sekvenciách, ktoré sú často celkovo veľmi odlišné. Pre dostatočné podobné sekvencie majú oba prístupy rovnakú efektivitu a kvalitu výsledku. Viacnásobne zarovnanie zahrňuje viac než dve sekvencie a využíva sa k identifikácii zakonzervovaných miest v skupine sekvencií alebo ku konštrukcii fylogenetického stromu, ktorý predstavuje vzájomné evolučné vzťahy medzi jednotlivými sekvenciami [24].

Pre vyhľadávanie podobných sekvencii sa využíva algoritmu BLAST (angl. Basic Local Alignmen Search Tool), ktorý porovnáva zadanú sekvenciu so skupinami alebo databázou sekvencií a identifikuje skupiny sekvencii, ktoré sú si podobné do určitej miery [25; 24].

Má viacero modifikácii:

i. blastn (angl. Nucleotide-nucleotide BLAST) – vyhľadá najviac podobné DNA sekvencie z DNA databáze k zadanej
ii. blastp (angl. Protein-protein BLAST) – vyhľadá najviac podobné sekvencie proteínu z proteínovej databáze k zadanej
iii. PSI-BLAST (angl. Position-Specific Iterative BLAST) – oproti blastp umožňuje nájsť i vzdialene príbuzné proteíny. Najskôr sa vytvorí zoznam blízkych príbuzných, tento výber je skombinovaný do všeobecného profilu sekvencie, ktorý zhrňuje významné vlastnosti použitých sekvencii. S týmto profilom je znova spustené vyhľadávanie, pri ktorom sa nájde väčšia skupina, z ktorej sa vytvorí širší profil a proces sa opakuje. S pripojením príbuzných proteínov do vyhľadávanie je PSI-BLAST oveľa viac citlivý na vyberanie vzdialených evolučných príbuzných ako štandardných blastp.

Rozšírenými nástrojmi pre vytváranie sekvenčných zarovnaní sú Clustal [26] a MUSCLE [27]. Clustal má dve verzie ClustalW pre príkazovú riadku a ClustalX s grafickým rozhraním.

2.2.3 Predikcia vplyvu mutácii na štruktúru a funkciu proteínov

V úvodných kapitolách bol už naznačený potenciálny prínos výpočtových metód pre predikciu vplyvu mutácií, buď už vzniknutých prirodzenou cestou, či umelým zásahom človeka, na funkciu proteínov. Použitie vhodných výpočtových metód môže priniesť veľkú finančnú a časovú úsporu vo výskume či medicínskej praxi a nie je preto divu, že bola za týmto účelom vyvinutá už celý rad rôznych nástrojov.

Nástroje pre predikciu vplyvu substitúcií vychádzajú zo sekvencie alebo štruktúry. Pozorovaním bolo zistené, že škodlivé mutácie sa pravdepodobnejšie vyskytujú na pozíciách, ktoré sú behom evolúcie konzervované v porovnaní s pozíciami, ktoré nie sú zakonzervované. To naznačuje potenciál použitia sekvencií pre predpovedanie vplyvu mutácii. Taktiež štruktúra môže byť použitá k predikcii, keďže chorobu spôsobujúce substitúcie majú spoločné štrukturálne vlastnosti, vyskytujú sa viac na povrchu proteínu. Využitie týchto vlastností viedlo k vzniku viacerých predikčných metód od množiny empirických pravidiel, cez skupinu strojovoučených techník obsahujúcich rozhodovacie stromy, s podporou vektorových strojov, neurálnych sieti a iné [28].

Okrem predpovedania, či je mutácia neutrálna, alebo nepriaznivo mení schopnosti proteínu (prípadne spôsobuje genetickú chorobu) dokážu predikčné metódy vrátiť aj presnosť predpovede, ale aj nárast alebo zníženie funkcionality, vplyv na štruktúru a iné atribúty.

3 Motivácia pre vývoj meta-servera

Za posledné desaťročie vzniklo mnoho rôznych metód na predikciu vplyvu mutácie na štruktúru a funkciu proteínov. Každá z nich pristupuje k predikciám iným spôsobom, využíva a analyzuje odlišné atribúty a prezentuje výsledky vlastným spôsobom. Samozrejme poskytujú tiež rôznu dôveryhodnosť a presnosť, k tomu využívajú rôzne dáta. Použitie viacerých principiálne odlišných metód pre riešenie určitého problému a kombinácia ich výsledkov môže značne navýšiť presnosť predikcií a poskytuje tiež predstavu o ich vierohodnosti. Avšak zadávanie výpočtov a vyplňovanie rovnakých dát do rôznych nástrojov, čakanie na výsledky z rôznych miest a porovnanie rôzne formátovaných výstupov predstavuje nemalú záťaž na užívateľa. Cieľom tohto projektu je vyvinúť meta-server, ktorý bude integrovať vybrané výpočtové nástroje, čím sa výrazne uľahčí práca vedcov pri výskume. Dosiahnutie tohto cieľa predpokladá zjednotenie vstupných dát, zobrazenie výsledkov na jednom mieste a tiež vytvorenie vlastnej konsenzuálnej predpovede, ktorá by vychádzala z výsledkov všetkých použitých metód.

4 Nástroje a možnosti predikcie

Všetky výpočtové metódy vychádzajú z určitých predpokladov, ku ktorým vedci a výskumníci došli experimentálnym pozorovaním. Tak bolo zistené, že škodlivé nsSNP majú sklony sa vyskytovať viac pod povrchom proteínu. Tie neutrálne skôr na povrchu. Ďalej sa tiež zistilo, že chorobné mutácie ovplyvňujú stabilitu molekuly. Spojením kritéria stability a štruktúry sa dá predikovať až 90% škodlivých mutácii. Zato neutrálnym mutáciám za použitia rovnakých kritérií vyhovuje iba 30%. Z toho vyplýva, že tieto pravidlá je možné použiť na rozoznanie škodlivých a neutrálnych mutácii [7].

Metódy založené na sekvenčnom kritériu vychádzajú z vedeckých pozorovaní, ktoré zistili, že škodlivé mutácie sa objavujú na miestach, ktoré sa počas evolúcie nemenili. Neutrálne sa vyskytujú skôr na premenlivých miestach sekvencie, čiže miestach, ktoré nie sú pre funkciu samotného proteínu až tak dôležité. Využívajú viacnásobného zarovnania, čím odlíši konzervované a nekonzervované miesta v sekvencii. Nasledujúca pravdepodobnosť sa ohodnotí podľa závažnosti zámeny aminokyseliny. Závisí to od podobnosti fyziochemických vlastností zamenenej aminokyseliny. Zo sekvencie a štruktúry je možné odvodiť celý rad ďalších kritérií užitočných pre predpoveď vplyvu mutácii na funkciu proteínov. Bližšie si ich priblížime pri charakteristike nástrojov využitých v meta-servery v nasledujúcich podkapitolách.

4.1 SIFT

SIFT (skratka z anglického Sorting Intolerant From Tolerant, roztriedenie netolerantných od tolerantných) je nástroj predikujúci vplyv mutácie na funkčnosť proteínu založenom na sekvenčnej zhode a fyzikálnych vlastnostiach aminokyselín [29; 30]. Môže byť aplikovaný na prirodzené mutácie, ale i na laboratórne vytvorené mutácie. Vychádza z predpokladu, že evolúcia proteínov má vzťah s ich funkciou. Čiže pozícia aminokyseliny dôležitá pre funkciu proteínu je nemenná v zarovnaní celej proteínovej rodiny, zatiaľ čo nedôležité pozície nesú rozdielne aminokyseliny v tomto zarovnaní. Prvá verzia vznikla už v roku 2001 a je vyvíjaný vo Výskumnom rakovinovom centre Freda Hutchisona v Seattly (orig. názov Fred Hutchinson Cancer Research Center, Seattle).

4.1.1.1 Funkčnosť

SIFT sa dá použiť dvoma spôsobmi. K rozhodnutiu, či je daná mutácia škodlivá a k ohodnoteniu tolerancie jednotlivých pozícii v sekvencii k mutácii.

Prevádza viackrokovú procedúru:

1. Vyhľadá podobné sekvencie pomocou PSI-BLAST (2.2.2) .
2. Vyberie blízko príbuzné sekvencie, ktoré zdieľajú stanovenú podobnosť a teda aj podobnú funkciu a štruktúru k dotazovanej.

3. Získa zarovnanie vybraných sekvencii z predchádzajúceho kroku. Alternatívou týchto troch krokov je zarovnanie zadané užívateľom.
4. Zo zarovnania spočíta pre každú pozíciu proteínu normalizované pravdepodobnosti pre všetky možné substitúcie.

V porovnaním s experimentálnymi dátami sa zistilo, že substitúcie s pravdepodobnosťou menšou než 0.05 sú často škodlivé. Tie väčšie alebo rovné ako 0.05 sú neutrálne.

4.1.1.2 Inštalácia

Na stránkach SIFTu (http://blocks.fhcrc.org/sift/) je možné využiť jeho on-line verziu alebo si stiahnuť skripty spustiteľne pod Unixom. Využijeme druhej možnosti, aby sme mohli používať nami vybranú databázovú sadu proteínových sekvencií. Zdrojové súbory sú dostupné na adrese http://blocks.fhcrc.org/sift/sift3.0.tar. K správnemu fungovaniu je nutný ešte nástroj BLAST. SIFT stačí rozbaliť a v konfiguračnom súbore nastaviť cesty k databázam a pomocným nástrojom. Viac informácii je k dispozícií v priložených súboroch k SIFTu.

4.1.1.3 Vstupy a výstupy

Pri využití on-line verzie SIFTu je možných viacero variant vstupov (sekvencia, dbSNP id a iné). My budeme využívať vlastnú lokálnu verziu, takže si popíše vstupy hlavne pre ňu. Nástroj SIFT vyžaduje na vstupe súbor so sekvenciou proteínu vo FASTA formáte, cestu k vybranej sekvenčnej databáze a súbor so substitúciami. Substitúcie sú zapísané vo formáte *XposY*, kde *X* je pôvodná aminokyselina na pozícii *pos* znamená aminokyselinou *Y*. Aminokyseliny sú zapísane jedným znakom tak, ako to definuje IUPAC. Na jednom riadku je povolená jedna substitúcia.

Výslednú predikciu SIFT vracia v jednom súbore s koncovkou `SIFTprediction`. K tomu vytvára súbor s viacnásobným zarovnaním, ktorý je možné ďalej spracovať inými bioinformatickými nástrojmi.

Tabuľka 1Ukážka výstupu SIFTu

Q10M	TOLERATED	0.12	2.71	22	98
Q11C	DELETERIOUS	0.04	2.74	23	98

Výstup obsahuje popis substitúcie, predpoveď jej škodlivosti a pravdepodobnosť výskytu takej substitúcie. Ďalším údajom vo výstupe je medián konzervovanosti, ktorým sa meria rozmanitosť sekvencii použitých v predikcii. Substitúcie označené ako chorobné pri mediáne väčšom ako 3,25 by mali byť brané z ohľadom na to, že predpoveď bola založená na blízko podobných sekvenciách. Takže pozícia substitúcie sa môže javiť ako zakonzervovaná čisto kvôli krátkemu času evolúcie, a nie vďaka jej dôležitosti pre funkciu a dôsledku toho môže byť nesprávne považovaná za chorobnú.

Ďalšie čísla udávajú počet použitých sekvencii na pozícii substitúcie (nepočítajú sa sekvencie s medzerou na tejto pozícii) a celkový počet sekvencií v zarovnaní.

4.2 MAPP

Predikčný nástroj MAPP (Multivariate Analysis of Protein Polymorphism, číže viacvariačná analýza proteínových polymorfizmov) vychádza z použitia viacerých na sebe nezávislých štatistických premenných [31]. Metóda sa neobmedzuje len na rozlíšenie mutácie na neutrálnu alebo škodlivú, ale stanovuje aj silu efektu od slabého až po závažný. Presnosť je závislá na podobnosti vstupných sekvencií a je najvyššia, keď sú k dispozícii odlišné orthologné sekvencie.

4.2.1.1 Funkčnosť

MAPP pozostáva so 7 krokov:

1. Zostaví sa viacnásobné zarovnanie orthologickych sekvencii, prípadne blízko príbuzných paralogických sekvencií, teda sekvencie proteínov, u ktorých sa neočakávajú významné rozdiely vo funkcii.
2. Na základe topológie a dĺžky vetiev fylogenetického stromu sa vypočíta váha jednotlivých sekvencií. Takže napr.: ak je v datasete niekoľko veľmi podobných sekvencií, váha každej z nich bude menšia, než u sekvencie, ktorá v datasete žiadnu blízko príbuznú nemá.
3. Pre každú pozíciu zarovnania sa vytvorí súhrn, v ktorom je každá z 20 možných aminokyselín reprezentovaná sumou váh sekvencií, ktoré na danej pozícii nesú danú aminokyselinu.
4. Tento súhrn je interpretovaný pomocou matice fyziko-chemických vlastností.
5. Výsledkom je vymedzenie vhodných fyziko-chemických vlastností aminokyselín pre každú pozíciu v podobe priemeru a variácie vlastností v sledovanom stĺpci zarovnania.
6. Odchýlky od stĺpca zarovnania sú získavané pre každú možnú variantu vypočítaním rozdielu jednotlivých vlastností medzi danou aminokyselinou a priemerom daného stĺpca.
7. Výsledkom je skóre, ktoré zhrňuje mieru odlišnosti danej aminokyseliny od fyziko-chemických vlastností vymedzených pre danú pozíciu zarovnania.

4.2.1.2 Inštalácia

Nástroj MAPP je možné stiahnuť na adrese http://mendel.stanford.edu/SidowLab/downloads/MAPP/ aj s podrobným návodom na použitie a testovacími dátami. Nástroj je napísaný v Jave a preto potrebuje k spusteniu nainštalovanú Javu 1.4 alebo vyššiu verziu. Vďaka jave je MAPP multiplatformný, čiže funguje na ľubovoľnom operačnom systéme. Na vstupe vyžaduje zarovnanie a fylogenetický strom príbuzných sekvencií, z toho dôvodu si vytvoríme pomocný skript, ktorý nám vytvorí súbor so zarovnaním a evolučným stromom homológov. Pre tieto účely sa využije nástroj

BLAST k vyhľadaní blízkych homológov v databázach, nástroj MUSCLE k vytvoreniu zarovnania a nástroj SEMPHY [32] k výpočtu evolučného stromu proteínov.

4.2.1.3 Vstupy a výstupy

MAPP vyžaduje na vstupe textový súbor so sekvenciou vo FASTA formáte. Názov tohto súboru sa špecifikuje voľbou `-f`. A textový súbor s evolučným stromom v zátvorkovej reprezentácii s dĺžkou vetiev špecifikovaný voľbou `-t`. Identifikátory sekvencií musia byť presne rovnaké a unikátne v oboch súboroch. MAPP je veľmi striktný na vstupné dáta.

MAPP vyprodukuje súbor s tabuľkou, v ktorej každý riadok pripadá na jednu pozíciu (stĺpec) v zarovnaní. Výstupný súbor je špecifikovaný voľbou `-o` inak sa výsledok vypíše na štandardný výstup. Pre každú pozíciu sú poskytnuté nasledujúce dáta:

1. pozícia v zarovnaní
2. mediánové MAPP skóre pre pozíciu
3. p-hodnota pre MAPP skóre
4. zarovnanie – aminokyseliny, ktoré sa v zarovnaní nachádzajú na danej pozícii
5. vážený podiel medzier na danej pozícii zarovnania
6. príznak indikujúci či pre danú pozíciu váha medzier prekračuje stanovenú hranicu
7. p-hodnota vyjadrujúca význam jednotlivých fyziko-chemických vlastností
8. MAPP skóre pre každú možnú aminokyselinovú zámenu na danej pozícii
9. p-hodnota každej zámeny, predpovedaný vplyv mutácie
10. zoznam neutrálnych mutácii pre danú pozíciu
11. zoznam škodlivých mutácii pre danú pozíciu

Výsledný súbor je jednoducho spracovateľný v MS Excel.

4.3 AUTO-MUTE

AUTO-MUTE je automatizovaný server na predikciu dôsledku ľudských nsSNP v kódujúcej oblasti sekvencie proteínov. Využíva kontrolovaný klasifikačný model založený na implementácii algoritmu Random Forest. K predpovedi využíva proteínovú štruktúru redukovanú do kolekcie bodov v 3-rozmernom priestore [33; 34].

4.3.1.1 Inštalácia

AUTO-MUTE má len webovú verziu na adrese http://proteins.gmu.edu/automute/AUTO-MUTE_nsSNPs.html, kde sa nachádza vstupný formulár. Pomôžeme si jednoduchým skriptom využívajúcim HTTP protokol na spojenie so serverom AUTO-MUTE. Požiadavkou POST mu zašle vstupné hodnoty a ako odpoveď získa stránku s HTML výsledkom, z ktorej sa vyparsruje predikcia a odfiltrujú sa dôležité výstupné dáta.

4.3.1.2 Vstupy a výstupy

AUTO-MUTE vracia okrem rozhodnutia o type polymorfizmu (chorobný alebo neutrálny)(Prediction) a pravdepodobnosti predikcie (Confid) aj ďalšie atribúty (Tabuľka 2).

Tabuľka 2 Ukážka výstupu AUTO-MUTU pre protein 1EGC a substitúciu I53T

PDB_ID	Chain	Mutation	Prediction	Confid	Vol	sT	Loc	Num	SS
1EGC	@	I53T	Disease	0.60	18.3	0.13	B	0	H

4.4 Zhrnutie a porovnanie nástrojov

Nástrojov a serverov na predikciu vplyvu mutácii na funkciu proteínov stále pribúda. Každoročne sa zvyšuje ich presnosť, mimo iné pokrytím, tzn. zdrojom a verziou použitej štruktúrnej a sekvenčnej databáze a celkovo tak rastie šanca na získanie správnych predikcií. Na druhú stranu s vyšším počtom nástrojov sa často aj zvyšuje obtiažnosť vyhodnotenia predpovedí pre vedcov, kde je koncový používateľ konfrontovaný so súborom predikovaných výsledkov, ktoré sú často vo vzájomnom nesúlade. Mnohé z nástrojov produkujú výsledky, ktoré sú ťažko pochopiteľné bez expertízy v danej oblasti. Vyžadujú rôzny stupeň skúsenosti, či už v informatike alebo biológii pri zadávaní výpočtu alebo vyhodnocovaní výsledkov [35]. Tieto nedostatky by mali byť brané na zreteľ pri návrhu meta-servera.

4.5 PBS

Nástroje pre predpovedanie vplyvu mutácii sú celkovo náročné na výpočtovú techniku. Využívajú nástroje ako napr. BLAST, ktorý prehľadáva proteínové databáze o veľkostiach niekoľko gigabajtov. Z toho je jasné, že výpočet potrvá istý čas, preto je dôležité rozumne rozvrhnúť výpočtový čas a systémové prostriedky pre jednotlivé predikcie. Jedným z možných riešení je využitie PBS (Portable Batch System). Tento nástroj je určený k prerozdeľovaniu výpočtových zdrojov úlohám – vykonáva plánovanie úloh. Primárne bol vyvinutý pre NASA začiatkom deväťdesiatych rokov. Má dve verzie open-source Torque PBS (predtým OpenPBS) a platenú PBS Pro. Umožňuje vyhradiť úlohám určené systémové prostriedky, výpočtový čas a podobne [36].

5 Meta-server – aplikácia

Našou úlohou je vytvoriť meta-server, ktorý po zadaní vstupných dát bude spúšťať viacero serverov a nástrojov a následne z nich zbierať výstupné dáta. Z výsledkov má vytvárať vlastnú konsenzuálnu predikciu, ktorá by mala byť vychádzajúc s predpovedí s rôznym prístupom presnejšia a dôveryhodnejšia. Prvá verzia meta-servera bude spolupracovať s predikčnými nástrojmi SIFT, AUTO-MUTE a MAPP, ktorými získame jednotlivé predikcie vplyvu mutácii na proteíny, presnosť predikcie a ďalšie výstupy, ktoré jednotlivé nástroje poskytujú. Ďalej budeme využívať skript z nástroja HotSpot Wizard pre získanie sekvencie zo štruktúry v PDB súbore a plánovací systém Torque PBS pre správne hospodárenie s výpočtovými prostriedkami.

HotSpot Wizard je nástroj pre proteínové inžinierstvo, vyvíjaný autormi Antonínom Pavelkom, Evou Chovancovou a Jiřím Damborským v Loschmidtových laboratóriách Masarykovej univerzity [37]. Tento nástroj slúži pre automatickú identifikáciu miest v proteíne modifikáciou, ktorých dôjde k zmene katalytických vlastností enzýmov. HotSpot Wizard integruje dohromady viacero bioinformatických databázy a výpočtových metód a vykonáva štruktúrne, funkčné a evolučné analýzy. Domovská stránka HotSpot Wizardu je na adrese http://loschmidt.chemi.muni.cz/hotspotwizard/.

Jednotlivé nástroje a programy sú implementované v rôznych programovacích jazykoch od skriptovacích až po kompilované. Samotný meta-server je implementovaný v PHP a na Unixové platformy z dôvodu, že používané nástroje sú primárne určené a vyvíjane na Unixových systémoch. Výsledkom spoločnej analýzy a návrhu s R. Wolným bolo rozdelenie tejto práce na dva projekty. Môj projekt sa zameriaval hlavne na spracovanie vstupov, plánovanie a spúšťanie výpočtov. Pre úplnosť bude spomenutá i časť s spracovaním výsledkov a vytváraním konsenzu, ktorej sa primárne venoval R. Wolný. Čitateľ tak bude mať úplnú predstavu o meta-servery.

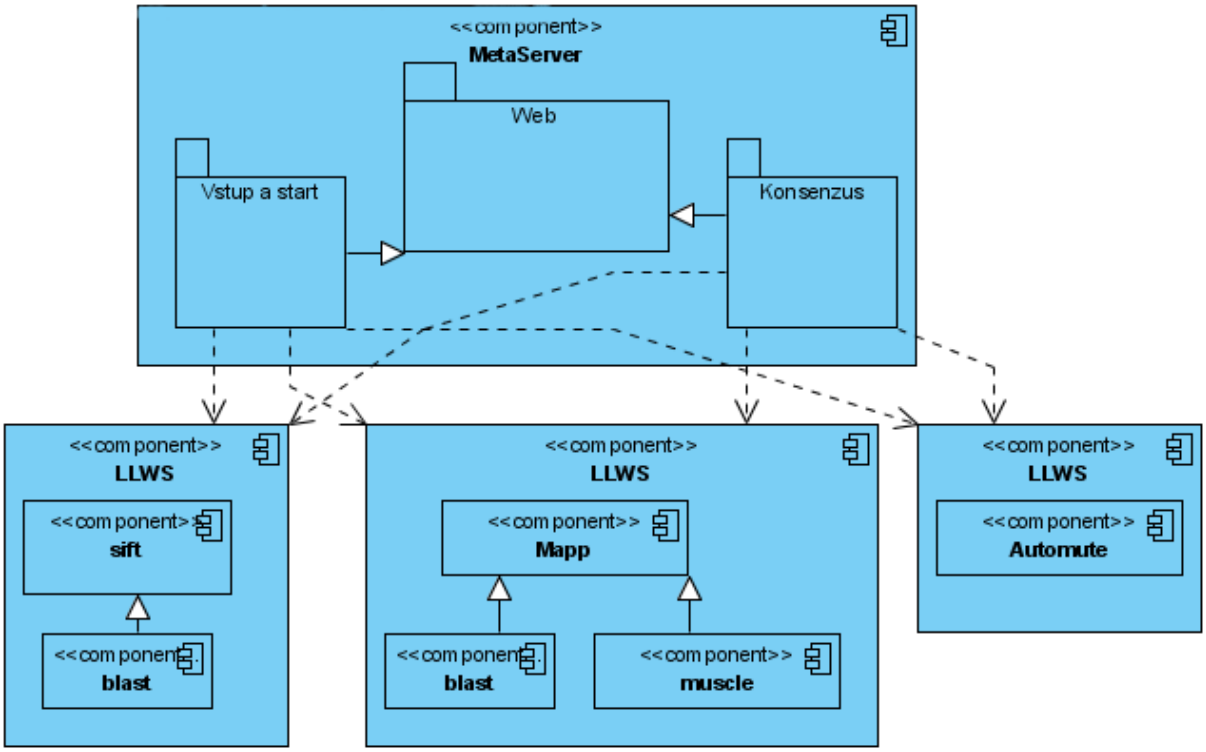
5.1 Analýza

Meta-server je možné rozdeliť na dve hlavné časti:

i. užívateľské rozhranie so vstupným formulárom
ii. stránka s výsledkom

Po odoslaní vstupného formulára sa spustí spracovanie náročných operácii na pozadí, ako je získavanie sekvencie zo štruktúry, predikcia pomocou jednotlivých nástrojov a spracovanie ich výsledkov. Ide teda hlavne o webovú aplikáciu implementovanú v jazyku PHP, ktorá spúšťa na pozadí ďalšie svoje časti. Takže interakcia v niektorých fázach práce s meta-serverom nebude okamžitá, ako by sa od webovej aplikácie očakávalo.

Meta-server bude rozdelený na komponenty, ktoré ho umožnia rozložiť na viacero strojov – serverov (Obrázok 4). Z obrázku vidno, že každý nástroj je schopný existovať na vlastnom servery, tak ako aj samotné webové rozhranie meta-servera, čím sa docieli rozloženie záťaže na výpočtovú techniku.



Obrázok 4 Rozdelenie meta-servera na komponenty

Mojou časťou meta-servera bol návrh tzv. vrstvy medzi predikčnými nástrojmi a meta-serverom pomenovanej LLWS (Loschmidt Laboratories Web Service) (5.1.3, 5.2.3), návrh obalových tried pre nástroje a časť ich samotnej implementácia (5.1.2, 5.2.2). Nasledujúce kapitoly sú venované hlavne týmto častiam.

5.1.1 Webové rozhranie

Webové rozhranie je časť meta-servera, ktorá interaguje so samotným užívateľom. Je to hlavne vstupný formulár, kde užívateľ zadá vstupy, pre ktoré chce vykonať predikciu, a výstupná HTML stránka, kde budú prehľadne vypísané výstupy jednotlivých nástrojov a súhrny konsenzuálny výsledok.

Keďže meta-server má spájať viacero predikčných nástrojov s rôznymi vstupmi (Obrázok 4), je nutné ich nejakým spôsobom zjednotiť do jedného vstupného formulára. Dôležitou položkou vstupného formulára je pre vstupy výber medzi predikciou zo sekvencie (vo FASTA formáte) alebo štruktúry (PDB id). Okrem zadania sekvencie, ktorú vyžadujú nástroje SIFT, MAPP a štruktúry,

ktorú vyžaduje AUTO-MUTE, bude povinným vstupom špecifikácia substitúcie. Užívateľ bude mať možnosť súčasne zadať väčší počet mutácii vo formáte *XposY* (4.1.1.3) oddelených prázdnymi znakmi, čiarkou alebo bodkočiarkou. Voliteľným vstupom bude súbor s viacnásobným zarovnaním, ktoré dokáže využiť MAPP.

Výsledková stránka bude obsahovať prehľadný výpis výsledkov jednotlivých použitých metód, prípadne výpis chybových hlásení, ku ktorým došlo počas behu jednotlivých nástrojov alebo celého meta-serveru. Najdôležitejšou časťou potom bude vypočítaný konsenzuálny výsledok. Výstupy jednotlivých nástrojov sa značne odlišujú, či už v terminológii, typoch výstupných parametrov, presnosti, pomocných atribútov a podobne. V niektorých výstupných dátach sa zhodujú, aj keď rozsah môže byť odlišný. Tieto dáta je preto nutné pretransformovať či už za účelom výpočtu konsenzu, alebo do prehľadu, kde si bude môcť užívateľ porovnať výsledky jednotlivých nástrojov. Po odoslaní formulára a po základnej kontrole vstupných dát sa spúšťa na pozadí vetva interne pomenovaná ako MSWS (Meta-Server WebService), ktorá následne vygeneruje ID úlohy a zobrazí ho užívateľovi.

5.1.2 MSWS

Táto časť má za úlohu vykonať všetky časovo náročnejšie operácie od odoslania formulára až po samotné spustenie jednotlivých nástrojov. Je to PHP skript spúšťaný z webového rozhrania na pozadí, aby nezdržoval vygenerovanie a zobrazenie ID úlohy.

Počas behu skriptu sa prekontrolujú vstupy pre každý spúšťaný nástroj zvlášť. Skontroluje sa správnosť jednotlivých substitúcii či už po syntaktickej, alebo sémantickej stránke. Starajú sa o to triedy navrhnuté pre každý nástroj a vychádzajúce zo spoločnej abstraktnej triedy. Tieto triedy sa postarajú aj o spustenie jednotlivých nástrojov, sprostredkovane pomocou špeciálne navrhnutého rozhrania pre spúšťanie bioinformatických nástrojov.

Nástroje sa samozrejme nespustia okamžite, ale zaradia sa do príslušných zásobníkov systému PBS, ktorý sa postará o samotné spustenie zadaného príkazu. Viac o tomto rozhraní v nasledujúcej časti.

5.1.3 LLWS

K spúšťaniu jednotlivých nástrojov a získavaniu ich výsledkov si navrhneme jednotné rozhranie, ktoré bude unifikovať volanie jednotlivých nástrojov. V podstate bude zaobaľovať konkrétny nástroj, malo by mať jednoduchú funkčnosť a byť univerzálna, čiže použiteľná aj pre iné účely, nie len pre náš meta-server. Pôjde teda o jednoduchú službu nastaviteľnú pomocou konfiguračných súborov pre jednotlivé nástroje.

Táto služba je vlastne vrstva nad jednotlivými nástrojmi s možnosťou distribuovatelnosti na viacerých serveroch, aby sa rozložila výpočtovú záťaž. Každý server bude mať svoju vlastnú LLWS

a každý nástroj na danom servery svoj vlastný konfiguračný súbor. Nový nástroj sa do služby LLWS pridá vytvorením konfiguračného súboru. Samozrejme musí spĺňať isté predpoklady, pretože ako už bolo spomínané a ešte veľakrát bude nástroje majú rôzne princípy funkčnosti, spôsob ukladania súborov s výsledkami a podobne. Preto pre takéto nástroje vytvoríme medzivrstvu (vrstvu medzi nástrojom a službou LLWS), ktorá sprístupní ovládanie nástroja cez LLWS.

Hlavnými funkciami LLWS je už okrem spomínaného spúšťania vybraného nástroja a získania jednotlivých výsledkov, tiež požiadavka, či už nástroj dokončil výpočet. Táto časť meta-servera je využiteľná aj v iných aplikácia a na iné účely. Bola preto v tomto duchu aj navrhovaná. Momentálne je už služba LLWS v praxi využívaná v projekte HotSpot Wizard pri práci s nástrojom Rate4Site [38], ktorý slúži pre výpočet stupňa konzervovanosti jednotlivých pozícii v proteíne.

5.1.3.1 Konfigurácia LLWS

Univerzálnosť služby budú zabezpečovať konfiguračné súbory, ktorých štruktúra by mala byť ľahko čitateľná, či už pre počítač, alebo človeka. K tomu si navrhneme syntax vychádzajúcu zo syntaxe INI konfiguračných súborov, ale tak aby stále splňovala pôvodne princípy INI a bola parsrovateľná dostupnými nástrojmi. Názvy súborov budú v tvare `nastroj.conf`, kde nástroj je názov nástroja, ku ktorému patrí a je case-sensitive. Tento názov sa bude ďalej používať aj k volaniu nástroja cez LLWS.

Z konfiguračného súboru sa LLWS dozvie akým príkazom spustí nástroj, aké vstupy má očakávať, a kde nájde súbory s výsledkami. Tiež obsahuje ďalšie interné informácie využívané LLWS, ako napríklad definované konštanty, k uľahčeniu zápisu konfiguračných súborov a zníženiu redundancie dát v ňom.

Syntax .conf súboru

Konfiguračný súbor je rozdelený do niekoľkých povinných a voliteľných blokov.

Blok `[command]`:

Obsahuje riadok s príkazom na spustenie nástroja aj s parametrami, ktoré sú v ďalších blokoch rozpísane. Parametre, ktoré sa nahradzujú vstupmi sú uvedené v tvare `$nazov_vstupu`. V príkaze sa tiež môžu vyskytnúť konštanty, ktorých sú ďalej nadefinované vo svojom bloku. Tie sa zapisujú v tvare `${nazov_konstanty}`. Špeciálne postavenie má konštanta `${output}`, ktorá má hodnotu vygenerovaného ID výpočtu (v konfiguračnom súbore sa nenastavuje jej hodnota v bloku konštánt).

Blok `[obligatory]`:

Tento blok obsahuje zoznam povinných vstupov, ktoré sa nahradzujú za parametre v spúšťacom príkaze v tvare `typ_vstupu = "nazov_vstupu1,nazov_vstupu2"`, kde typ vstupu je buď `File` alebo `Data`. Na ďalších riadkoch môžu byť uvedené defaultné hodnoty pre jednotlivé vstupy, ak nie sú zadané.
`Nazov_vstupu = "defaultna_hodnota"`.

Blok `[optional]`:

Totožný s blokom povinných parametrov s tým rozdielom, že obsahuje nepovinné vstupy. Ak nemá nepovinný vstup nastavenú defaultnú hodnotu odmaže sa zo spúšťaného príkazu.

Blok `[check]`:

Blok s adresou k súboru, podľa ktorého sa rozpozná koniec behu nástroja. Môžu sa v ňom vyskytnúť konštanty a je zapísaný v tvare: `check = "cesta_k_suboru"`.

Blok `[results]`:

V tomto bloku definujeme cesty k súborom s výsledkami. Taktiež môže obsahovať konštanty. `Oznacenie_subora = "cesta_k_vyslednemu_suboru"`, kde oznacenie_subora budeme používať v požiadavku pre vrátenie daného výsledného súboru. Ak nástroj spúšťaný LLWS službou vytvára jeden súbor s výsledkom, ktorý zároveň slúži ako kontrolný, nemusí byť tento blok uvedený.

Blok `[const]`:

Blok definovaných konštánt platiacich v rámci konfiguračného súboru. Konštanty sú definované v tvare `nazov_konstanty = "hodnota_konstanty"`.

Na poradí jednotlivých blokov nezáleží, ale je odporúčané dodržať poradie vyššie uvedené poradie kvôli prehľadnosti.

Príklad .conf súboru, kde výsledky sú označené ako `R1`, `R2` a `R3`, súbor `check` je kontrolný (slúži na zistenie ukončenia výpočtu):

```
[command]
command="perl  ${path}/consurf_v0.95.pl ${output} $query $evalue $max_hits_nb $msa
$target_msa $tree"
[obligatory]
File = "query,msa,tree"
Data = "evalue,max_hits_nb,target_msa"
[check]
check = " ${path}/${output}/finished"
[results]
```

```
R1=“${path} /${output}/MSA.fas“
R2=“${path}/${output}/conservation.txt“
R3=“${path}/${output}/tree.nwk“
[const]
path = “ /usr/lib/aas/sw/tools/consurf“
```

Príklad .conf súboru, kde výsledok je v jednom súbore (`check`), ktorý je zároveň kontrolný súbor:

```
[command]
command=“/usr/aas/sw/tools/sift/sift3.0/bin/SIFT_for_submitting_fasta_seq.csh
$fasta /usr/aas/sw/tools/blast/blast-2.2.18/db/FASTA/uniprot_sprot.fasta $subst“
[obligatory]
File = “fasta,subst“
[check]
check=“/usr/aas/sw/tools/sift/sift3.0/tmp/${output}.SIFTprediction“
```

5.1.4 Konsenzus

Ako už bolo spomenuté beh jednotlivých nástrojov je časovo náročný a vopred sa nedá odhadnúť doba trvania ich výpočtu. Preto k prebratiu výsledkov si navrhneme skript, ktorý sa počas svojho behu jednorázovo pokúsi získať výsledky. K tomu opäť využije služby LLWS. Celý skript bude periodicky spúšťaný systémovým nástrojom cron v istých intervaloch.

Skript počas jedného behu bude obstarávať všetky nedokončené úlohy. Ak získa predpovede od všetkých nástrojov k danej úlohe, ďalej ich spracuje a predpripraví k výpočtu konsenzu a vytvoreniu výstupnej stránky. Tak isto ako pri skripte MSWS sa o získavanie výsledkov od jednotlivých nástrojov, spracovanie a predpripravenie starajú metódy z tried zapuzdrujúcich jednotlivé nástroje.

Ku vlastnému konsenzuálnemu výsledku dospejeme určitým výpočtom, v ktorom vstupmi sú výsledky z použitých metód. Výpočet konsenzu sa prevádza za účelom zvýšenia presnosti predpovede. Najjednoduchší výpočet je klasický vážený priemer.

$p = (a_1p_1 + a_2p_2 + a_3p_3 + a_4p_4 + \ldots + a_np_n)/n$

p_i … predpoveď (-1 = škodlivý, 1 = neutrálny)

a_i … odhadovaná pravdepodobnosť, že predpoveď p_i je správna

Každý nástroj v tomto konsenze má rovnakú váhu. Takže aj menej presné metódy majú rovnaký vplyv na výsledný konsenzus ako tie presnejšie. Ďalšou možnosťou je uprednostniť niektoré nástroje tým, že im dáme väčšiu váhu na ovplyvnenie výsledku. Jednotlivé nástroje ohodnotíme podľa ich presnosti predpovede v porovnaní s experimentálne overenými mutáciami.

Konsenzus nášho meta-servera sa bude počítať v niekoľkých krokoch:

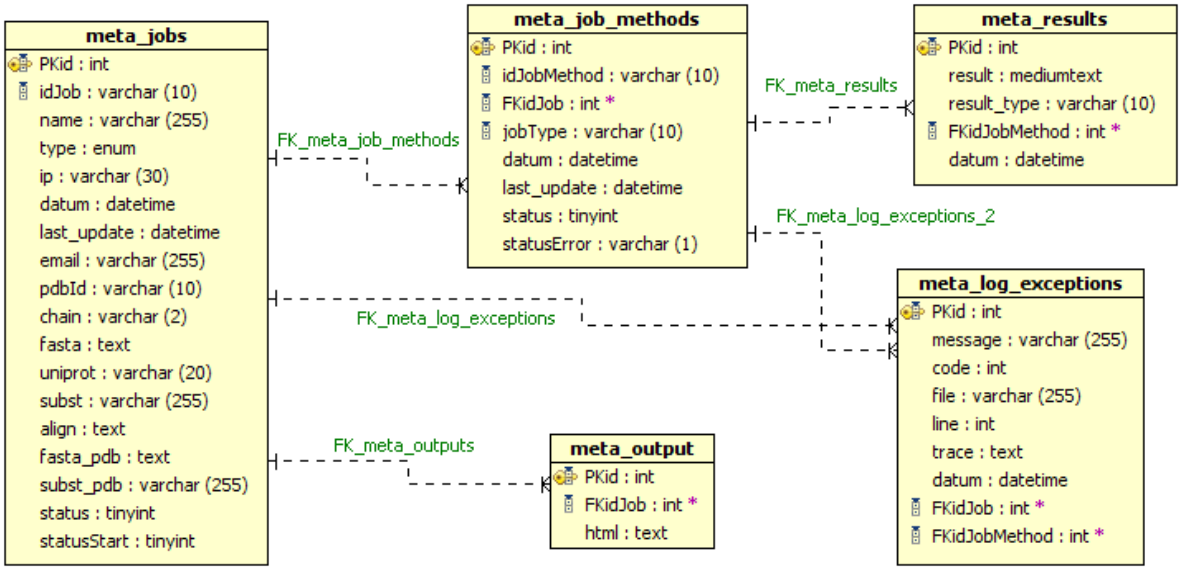
1. Pre každý nástroj si určíme hodnotu predikcie *HP* (-1 škodlivý, 1 neutrálny) a dôverihodnosť predikcie *Conf*.

a. Automute priamo udáva dôverihodnosť
b. pre MAPP pôužijeme ako dôverihodnosť *p-hodnotu*, kde 0,00 je 100% a 1,00 je 0% (napr.: 0.05 => 95%)
c. pre SIFT vypočítame dôveryhodnosť zo *SIFT score*
    - Pre škodlivú to bude *SIFT score* 0 – 0.05 na 100 - 0% dôveryhodnosti
    - pre neutrálnu to bude *SIFT score* 0.05 – 1 na 0 – 100% dôveryhodnosti

2. Hodnotu konsenzu (*consensual value*) počítame ako súčet súčinov *HP* a *Conf* pre každý nástroj, ktorý dáva výsledok. Ak je výsledok menší nule je mutácia predikovaná ako škodlivá, inak ako neutrálna.
3. Okrem predikcie konsenzus počíta aj *tools conformity* a *average conficence*.
    a. *tools conformity* je podiel počtu nástrojov, ktoré predikovali zhodne s konsenzuálnym výsledkom a počtu všetkých nástrojov, ktoré dali nejakú predikciu.
    b. *average confidence* udáva priemernú dôveryhodnosť všetkých nástrojov, ktoré dali predikciu bez ohľadu na výslednú predikciu.

5.1.5 Databáza pre meta-server

K uchovaniu vstupných dát, informácii o jednotlivých úlohách a ich podúlohách a prenosu dáta medzi jednotlivými časťami bude meta-server využívať databáze. Vzťahy medzi jednotlivými tabuľkami sú zobrazené na obrázku (Obrázok 5).



Obrázok 5 ER-diagram meta-servera

Tabuľka `meta_jobs` obsahuje základne informácie o úlohe a vstupné dáta, je prepojená väzbou 1:N s `meta_job_methods` s údajmi o jednotlivých podúlohách. S tabuľkou `meta_job_methods` sa spája `meta_results` väzbou 1:N. Obsahuje výsledky jednotlivých nástrojov. Poslednou tabuľkou je `meta_output`, ktorá je vo vzťahu s `meta_jobs` 1:1 a uchováva

časť výslednej HTML stránky s výsledkami jednotlivých nástrojov a konsenzus. Databáza bude implementovaná na databázovom systéme MySQL.

5.2 Implementácia

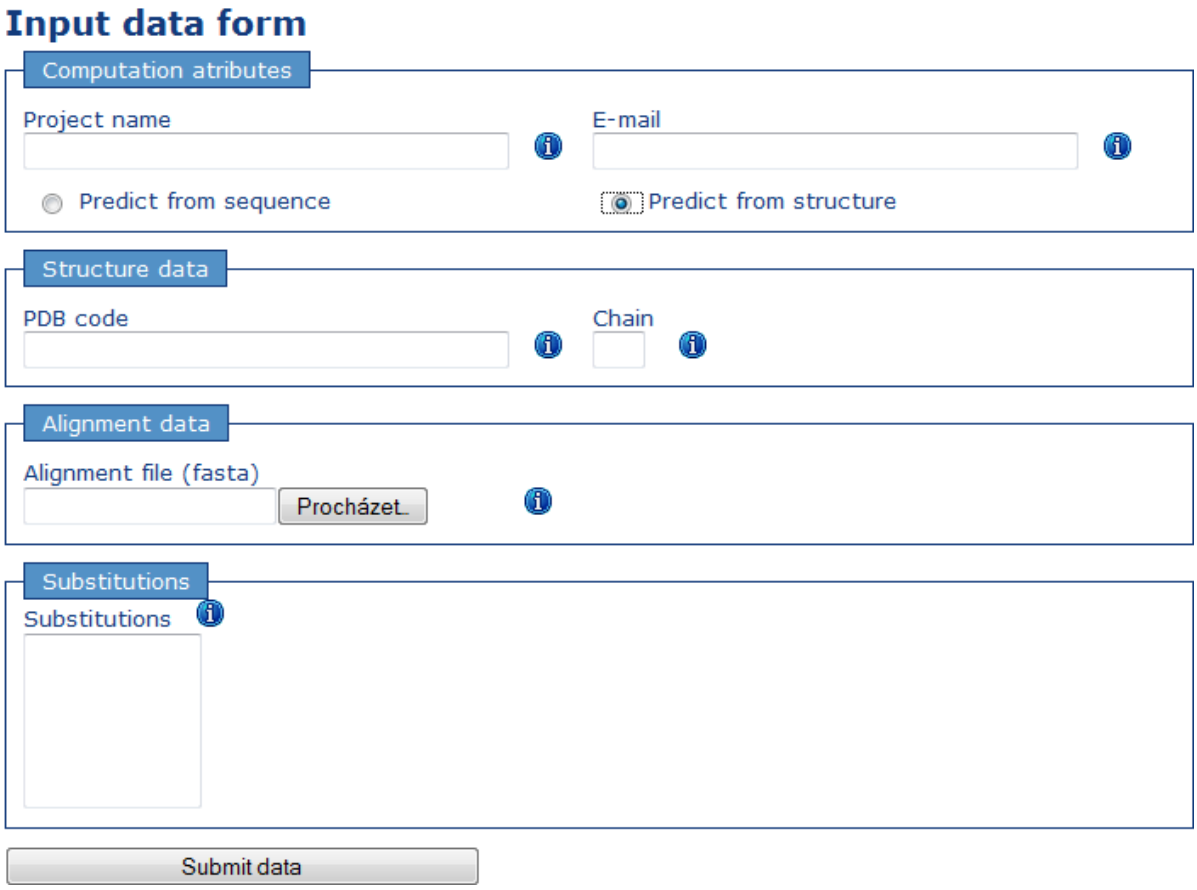
Meta-server je implementovaný v jazyku PHP 5, ktorý poskytuje dobrú podporu objektovo-orientovaného programovania. Každý predikčný nástroj je zapuzdrený vo vlastnom objekte, ktorého metódy obstarávajú všetko od načítania dát pre konkrétny nástroj až po prezentáciu výsledku. Tieto objekty vychádzajú zo spoločného abstraktného objektu tak, aby sa vytvorila unifikovanosť pre inak rozdielne predikčné nástroje. O chod meta-servera sa starajú tri hlavné nosne obsluhujúce triedy (handlery), `inputHandler` príjma vstupy, `proceedHandler` spracuje vstupy a `outputHandler` získava a vyhodnocuje výsledky. Bližšie si všetky objekty popíšeme v nasledujúcich častiach rozdelených tak ako pri návrhu meta-servera aj s konkrétnymi objektmi a ich metódami patriacimi k jednotlivým častiam.

5.2.1 Webové rozhranie

Webový formulár (Obrázok 6), stránka s ID úlohy a s výsledkom sú zabudované do redakčného systému Expo|CMS [39]. Pre tento systém sme sa rozhodli z časových dôvodov, pretože to bol jediný redakčný systém, s ktorým mal jeden z nás spoluriešiteľov dlhodobejšie skúsenosti a poznal jeho mechanizmy na dobrej úrovni. Meta-server samotný nie je závislý na tomto systéme a po malých úpravách a prispôsobeniach môže byť vstavaný do hocijakého redakčného systému resp. fungovať aj ako samostatná aplikácia.

Trieda `inputHandler`

Po odoslaní formulára sa objekt `inpuHandler` postará o jeho načítanie a skontroluje vyplnenie povinných údajov (pomocou metódy `checkAndLoadInputData`). Samozrejme má za úlohu taktiež prideliť úlohe unikátnu identifikačnú hodnotu (`generateId`). Všetky tieto údaje týkajúce sa požiadavku na spustenie úlohy (jej ID, vstupne hodnoty, spúšťane nástroje a iné) sa uložia do databáze pre potreby ďalších častí meta-servera. V poslednej fázy sa predá úloha skriptu MSWS (`startComputation`), ktorý ďalej pokračuje v spracúvaní požiadavky na pozadí a užívateľovi sa zobrazí priradené ID úlohy s odkazom, kde po skončení úlohy nájde výsledok.



Obrázok 6 Ukážka vstupného formulára meta-servera

5.2.2 MSWS

MSWS je jednoduchý skript riedený objektom `proceedHandler`, ktorý dostáva ako vstupný parameter ID úlohy. Pomocou neho si z databáze vytiahne všetky potrebné dáta k spracúvanej úlohe.

Trieda `proceedHandler`

Na začiatku si vytiahne potrebné dáta z databáze a inicializuje objekty obsluhujúce jednotlivé predikčné nástroje (`loadInputsFromDb`, `loadMethodsFromDb`). Po inicializovaní sa týmto objektom nástrojov predajú vstupné hodnoty, pričom každý objekt tieto hodnoty skontroluje podľa požiadaviek konkrétneho nástroja (`checkAndLoadInputToMethods`). Na záver sa službou LLWS spustia jednotlivé nástroje a do databáze sa uložia ID hodnoty k podúlohám celej úlohy. Podúlohy sú jednotlivé predikčné nástroje spustené v rámci danej úlohy.

Pozn.: V skutočnosti sa tieto nástroje nespustia, len sa odošle požiadavka do LLWS, ktorá príkaz na spustenie zaradí do zásobníka v PBS a vráti vygenerované ID výpočtu.

Abstraktná trieda `methodSNP.abstract`

Abstraktná trieda `methodSNP.abstract` vychádza z interfacu `methodSNP.interface`, ktorý obsahuje základné metódy, ktoré musia obsahovať ostatné triedy z neho odvodené. Sú nimi `loadInputData`, ktorá načíta a skontroluje vstupné hodnoty a `startMethod`, ktorá posiela požiadavku na LLWS k spusteniu výpočtu. Obsahuje ešte ďalšie metódy, ktoré sa využívajú v triede `outputHandler` pri spracovaní výsledku (napr.: `parseOutputs`, `takeOverOutput`). V abstraktnej triede sú nadefinované všetky spoločné metódy pre koncovo odvodené triedy nástrojov.

5.2.3 LLWS

Služba LLWS je funkčne samostatná časť meta-servera využiteľná aj v iných aplikáciách. Jej funkčnosť je veľmi jednoduchá a umožňuje rozložiť výpočty na viacero serverov. To znamená, že každý predikčný nástroj môže byť na inom servery a teda každý musí obsahovať službu LLWS s príslušným konfiguračným súborom k danému nástroju. Je to opäť jednoduchý skript s tromi funkciami riadený objektom `method`. Prijíma požiadavku na spustenie výpočtu, požiadavka na kontrolu ukončenia výpočtu a vrátenie súboru s výsledkom. V prvých dvoch prípadoch vracia dáta vo formáte INI konfiguračných súborov, v treťom prípade vracia obsah výsledného súboru.

Trieda `method`

Základnou metódou objektu je `loadConfigFile`, ktorá pri prijatí požiadavku na štart nástroja načíta konfiguračný súbor konkrétneho nástroja. Z neho si zistí aké vstupy nástroj očakáva, kam uloží výsledky a akým príkazom sa spustí. Následne načíta vstupy (`loadDataFromRequest`), keďže je to jednoduchá služba tak očakáva, že všetky vstupy sú v poriadku, takže predpokladá, že o kontrolu sa postaral užívateľ resp. aplikácia volajúca službu LLWS. Záverom sa prevedie príkaz z konfiguračného súboru. V prípade meta-servera sa príkaz zaradí do zásobníka v PBS systéme. Vygeneruje ID podúlohy a vráti ho volajúcemu.

Pri kontrole ukončenia výpočtu LLWS vyžaduje ID konkrétnej podúlohy. Metódou `loadResultFile` si načíta cestu ku kontrolnému súboru a následne metódou `checkResult` skontroluje jeho existenciu. Potom vráti `status=yes` pri ukončení, inak `status=no`.

Tretia funkcia LLWS je principiálne rovnaká až na to že nevracia stav predikcie, ale rovno obsah požadovaného súboru. Ak existuje viac výsledných súborov danej podúlohy, súbor nášho záujmu musí byť v požiadavku špecifikovaný.

Postup práce s LLWS:

1. Zavolá sa služba LLWS s parametrom metódy a akcie `start`
   `http://url_llws/llws.php?method=[method]&action=start`
   `url_llws` – URL adresa LLWS služby

`[method]` – štandardizovaný case-sensitive názov vybraného nástroja

Spolu s volaním LLWS služby sa posielajú HTTP metódu POST alebo GET aj príslušné vstupy, ktoré daný nástroj podľa svojho konfiguračného súboru vyžaduje.

Príklad spustenia nástroja SIFT v meta-servery:

```
http://localhost/llws/xlisak00/llws.php?method=SIFT&action=start
```

Pozn.: meta-server i služba LLWS sa nachádza na rovnakom servery.

2. LLWS vyberie a načíta potrebný .conf súbor. So súboru zistí aký príkaz spúšťa, aké vstupy má očakávať, a kde nájde výsledky výpočtu spusteného nástroja.

   Pozn.: Adresárová štruktúra s .conf súbormi pre jednotlivé nástroje je `conf/[method]/[method].conf`, kde `[method]` je štandardizovaný case-sensitive názov vybraného nástroja

3. LLWS upraví príkaz na spustenie nástroja (pridá unixové parametre pre spustenie na pozadí, vloží zadané vstupy, odstráni nezadané voliteľné parametre), spustí príkaz, vytvorí súbor, ktorý vyplní cestami k výsledkom a kontrolnému súboru a vráti vygenerované ID výpočtu.

4. Na zistenie ukončenia výpočtu sa volá služba LLWS s parametrom metódy, ID výpočtu a akcie check

   ```
   http://url_llws/
   llws.php?method=[method]&idJob=[id_job]&action=check
   ```

   `url_llws` – URL adresa LLWS služby

   `[method]` – štandardizovaný case-sensitive názov vybraného nástroja

   `[id_job]` – ID výpočtu

   Pri ukončení vracia `status=yes`, inak `status=no`. Ukončenie závisí na existencii kontrolného súboru (cesta k nemu je v súbore vytvorenom pri spúšťaní výpočtu).

5. Požiadavka na vrátenie konkrétneho súboru s výsledkom sa opäť robí volaním služby LLWS s s parametrom metódy, ID výpočtu, akcie `get` a nepovinným súbor.

   ```
   http://url_llws/llws.php?method=[
   method]&idJob=[id_job]&action=get&file=[file]
   ```

   `url_llws` – URL adresa LLWS služby

   `[method]` – štandardizovaný case-sensitive názov vybraného nástroja

`[id_job]` – ID výpočtu

`[file]` – interné označenia súboru s výsledkami, ak nie je zadané implicitne vracia súbor `check`

Pozn.: `check` je kontrolný súbor, v niektorých prípadoch sa môže jednať zároveň o súbor s výsledkom výpočtu (pozri syntax .conf súboru).

5.2.4 Konsenzus

Záverečnou fázou práce meta-servera je zber a spracovanie výsledkov. O to sa stará posledný z troch handlerov objekt `outputHandler`. Tento objekt tak isto ako `proceedHandler` na začiatku inicializuje metódou `loadMethodsFromDb` jednotlivé použité nástroje. Využitím ich metódy `takeOverOutput` získame cez službu LLWS súbor s výsledkom a metódou `parseOutputs` ho rozparsrujeme a odfiltrujeme si dáta potrebné pre náš konsenzus. O výpočet samotného konsenzu sa stará inštancia triedy `consensualPrediction`. Výsledná stránka s predpoveďou (Obrázok 7) je po častiach generovaná metódou `createHTMLOutput` inštancie konsenzu a metódami `createHTMLFromOutputs` inštancií tried jednotlivých predikčných nástrojov. Celý tento postup sa deje až po ukončení všetkých výpočtov konkrétnej úlohy.

Project summary

Project name	**Muj projekt**
Date of submition	2009-04-10 13:55:59
Project code	KlBA32BK8f
E-mail	
PDB id	2PSE
Chain	A
Sequence	>2PSE KVYDPEQRKRMITGPQWWARCKQMNVLDSFINYYDSEKHAENAVIFLHGNATSSYLWRHV VPHIEPVARCIIPDLIGMGKSGKSGNGSYRLLDHYKYLTAWFELLNLPKKIIFVGHDWGA ALAFHYAYEHQDRIKAIVHMESVVDVIESWDEWPDIEEDIALIKSEEGEKMVLENNFFVE TVLPSKIMRKLEPEEFAAYLEPFKEKGEVRRPTLSWPREIPLVKGGKPDVVQIVRNYNAY LRASDDLPKLFIESDPGFFSNAIVEGAKKFPNTEFVKVKGLHFLQEDAPDEMGKYIKSFV ERVLK
Substitutions	F33R I34M E44G A54G
Reindexed subst.	F30R I31M E41G A51G

Automute

PDB ID	Chain	Mutation	Prediction	Confid	Vol	sT	Loc	Num	SS
2PSE	A	F33R	Disease	0.57	11.9	0.13	B	0	S
2PSE	A	I34M	Neutral	0.61	18.5	0.11	B	0	C
2PSE	A	E44G	Neutral	0.60	15.3	0.19	S	5	C
2PSE	A	A54G	Disease	0.70	22.7	0.18	B	0	C

MAPP

Substitution	Column score	Column p-value	Prediction	Gap weight	Hydropathy	Polarity	Charge	Volume	Free energy alpha	Free energy beta
F33R(F30R)	23.71	0.000087	Disease	0.0000	0.0266	0.0730	0.9999	0.1615	0.1093	0.1961
I34M(I31M)	21.44	0.000155	Disease	0.0000	0.0503	0.0229	0.0166	0.1063	0.5890	0.0258
E44G(E41G)	23.22	0.000098	Disease	0.0000	0.0392	0.0245	0.0173	0.1353	0.4531	0.3427
A54G(A51G)	23.34	0.000095	Disease	0.0000	0.0633	0.5697	0.0505	0.6850	0.9796	0.9994

COMPLETE MAPP RESULTS FOR DOWNLOAD

SIFT

Substitution	Prediction	SIFT score	Median	xxx	xxxx
F33R(F30R)	Neutral	1.00	2.87	16	16
I34M(I31M)	Neutral	1.00	2.87	16	16
E44G(E41G)	Neutral	1.00	3.11	10	16
A54G(A51G)	Disease	0.04	2.87	16	16

Consensual prediction

Substitution	Consesual value	Tools included	Prediction
F33R	0.56991335	AUTOMUTE MAPP SIFT	Disease
I34M	-0.6101552	AUTOMUTE MAPP SIFT	Neutral
E44G	-0.60009786	AUTOMUTE MAPP SIFT	Neutral
A54G	1.89990503	AUTOMUTE MAPP SIFT	Disease

Obrázok 7 Ukážka výstupu meta-servera

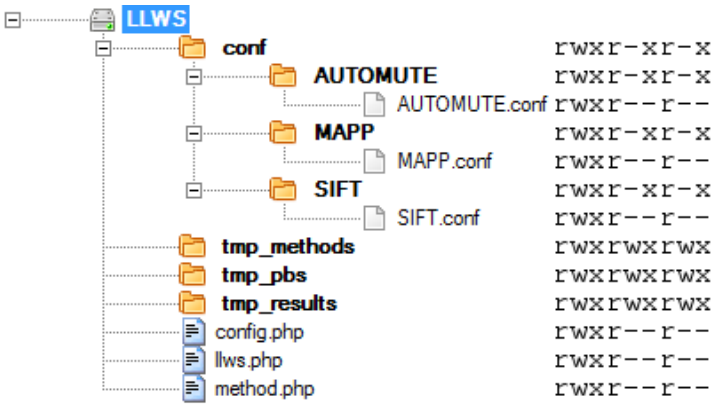
5.3 Inštalácia meta-servera

Meta-server je stavaný pre unixovo založené systémy. Bol testovaný a sprevádzkovaný na systéme Debian, servery Apache 2.2.3 s podporou PHP 5.2 a databázy MySQL 5.0.

LLWS

Pre každý nástroj, ktorý chceme spúšťať pomocou služby LLWS, si vytvoríme konfiguračný súbor (5.1.3.1) uložený v `./conf/nazov_nastroja/nazov_nastroja.conf`, kde `nazov_nastroja` je vhodne zvolené označenie nástroja, podľa ktorého budeme neskôr nástroj volať. Ostatné skripty

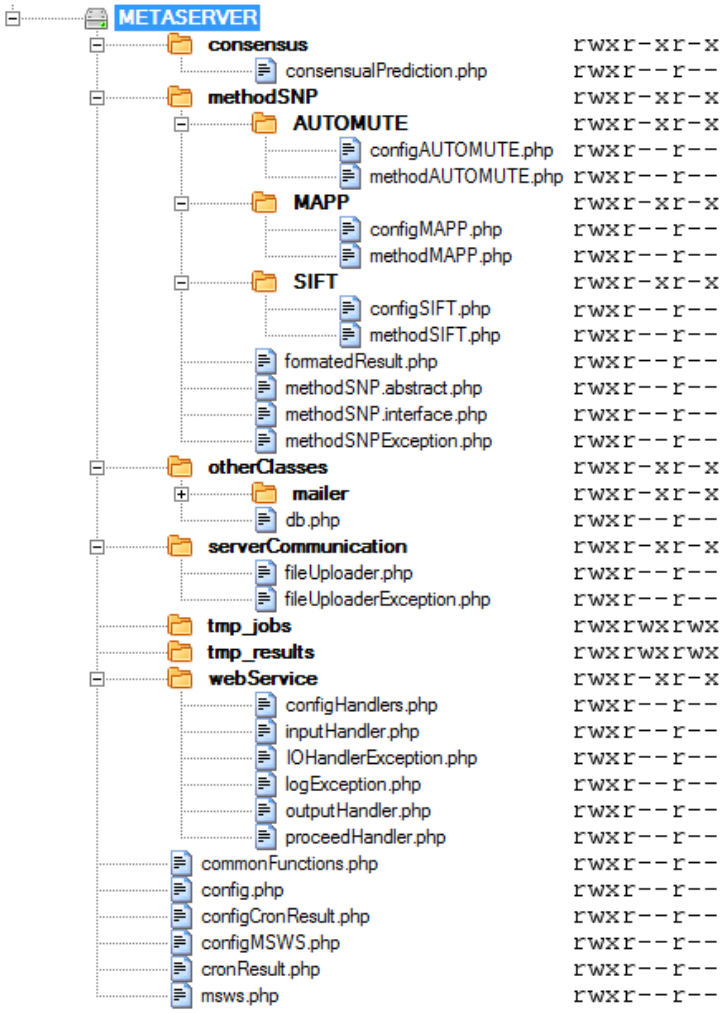
a adresáre služby LLWS nakopírujeme na server s prístupovými právami ako na obrázku dole (Obrázok 8).



Obrázok 8 Adresárová štruktúra služby LLWS

Meta-server

Adresárovú štruktúru a prístupové práva skriptov meta-servera je treba nastaviť podľa obrázka Obrázok 9. V konfiguračných súboroch meta-servera (`config.php`, `configCronResult.php`, `configMSWS.php`) nastavíme prístupy k databáze. V `configCronResult.php` nastavíme cesty ako napr. k skriptu extrahujúcemu sekvenciu z PDB súboru. V konfiguračných súborov objektov zapuzdrujúcich predikčné nástroje nastavíme cesty k ich LLWS službe. Ak chceme doplniť triedu pre nový nástroj do meta-servera, musí byť odvodená od `method_abstract` a byť uložená v `methodSNP/Nazov_nastroja/methodNazov_nastroja.php`, kde `Nazov_nastroja` je jeho nami zvolené označenie.



Obrázok 9 Adresárová štruktúra meta-servera

5.4 Testovanie na reálnych dátach

V záverečnej časti práce sa budeme venovať testovaniu meta-servera na reálnych dátach. Výsledky porovnáme so sadami premutovaných proteínov, zrovnáme výsledky nástrojov medzi sebou a konsenzuálnym výsledkom a prípadne navrhneme možné spôsoby jeho vylepšenia do budúcna. Pri interpretácii výsledkov treba pamätať, že ich porovnanie s datasetmi nie je jednoduché, pretože nie sú unifikovane. To znamená, že aj keď boli získavané experimentálne, či už *in vivo* alebo *in vitro* alebo pri klinických štúdiách, tak na posúdenie vplyvu mutácie boli použité rozdielne kritéria. Jednotlivé pozície v sekvencii konkrétneho proteínu premutujú niekoľkými aminokyselinami a výsledný vplyv sa rozradí do jednej z kategórii (tie sú väčšinou tri alebo štyri podľa vážnosti vplyvu a sú označované ako od najmenej vážnych: + > +- > -+ > -). Tieto kategórie pre každý proteín zvlášť zlúčime do dvoch kategórií. Hranicu zvolíme tak, aby presnosť nástrojov bola maximálna.

K dispozícii k testovaniu sme mali datasety pre:

i. lysozým [40], PDB kód 1LYD
ii. HIV proteázu [41], PDB kód 1AJV

iii. reverznú transkriptázu [42; 43], PDB kód 3HVT.

Spočítame priemerné presnosti nástrojov a konsenzov. Priemerná presnosť konsenzuálnej predikcie meta-servera pri zahrnutí všetkých troch nástrojov bola 73,80 %, presnosť jednotlivých nástrojov potom 61,38 % pre AUTO-MUTE, 69,78 % pre MAPP, 73,63 % pre SIFT. V tabuľke Tabuľka 3 sú uvedené jednotlivé miery presnosti vypočítané pre konsenzus a každú metódu zvlášť.

Tabuľka 3 Presnosť predikcií

3HVT	**AUTO-MUTE**	**MAPP**	**SIFT**	**Cons3[5]**	**Cons2AS[6]**	**Cons2MS[7]**	**Cons2AM[8]**
Accuracy[1]	66,39 %	61,39 %	76,92 %	75,00 %	76,10 %	77,75 %	59,89 %
True positive rate [2]	72,22 %	50,78 %	88,67 %	80,86 %	91,80 %	83,98 %	53,91 %
True negative rate[3]	52,78 %	87,50 %	49,07 %	61,11 %	38,89 %	62,96 %	74,07 %
Precision[4]	78,11 %	90,91 %	80,50 %	83,13 %	78,07 %	84,31 %	83,13 %

1AJV	**AUTO-MUTE**	**MAPP**	**SIFT**	**Cons3**	**Cons2AS**	**Cons2MS**	**Cons2AM**
Accuracy[1]	62,65 %	77,38 %	77,08 %	78,27 %	80,95 %	79,17 %	76,49 %
True positive rate [2]	66,82 %	92,44 %	77,78 %	94,67 %	87,11 %	96,00 %	92,44 %
True negative rate[3]	54,21 %	46,85 %	75,68 %	45,05 %	68,47 %	45,05 %	44,14 %
Precision[4]	74,74 %	77,90 %	86,63 %	77,74 %	84,85 %	77,98 %	77,04 %

1LYD	**AUTO-MUTE**	**MAPP**	**SIFT**	**Cons3**	**Cons2AS**	**Cons2MS**	**Cons2AM**
Accuracy[1]	55,09 %	70,57 %	66,90 %	68,14 %	61,19 %	68,98 %	68,93 %
True positive rate [2]	71,16 %	72,26 %	84,33 %	80,09 %	88,40 %	79,78 %	73,35 %
True negative rate[3]	47,58 %	69,79 %	58,82 %	62,60 %	48,58 %	63,98 %	66,88 %
Precision[4]	38,84 %	52,57 %	48,69 %	49,81 %	44,34 %	50,65 %	50,65 %

ALL	**AUTO-MUTE**	**MAPP**	**SIFT**	**Cons3**	**Cons2AS**	**Cons2MS**	**Cons2AM**
Accuracy[1]	61,38 %	69,78 %	73,63 %	73,80 %	72,75 %	75,30 %	68,44 %
True positive rate [2]	70,07 %	71,83 %	83,59 %	85,21 %	89,10 %	86,59 %	73,23 %
True negative rate[3]	51,52 %	68,05 %	61,19 %	56,25 %	51,98 %	57,33 %	61,70 %
Precision[4]	63,90 %	73,79 %	71,94 %	70,22 %	69,09 %	70,98 %	70,27 %

[1] udáva celkovú presnosť predikcie, teda pomer mutácií, ktoré boli predikované správne ku celkovému počtu mutácií

[2] udáva pomer správne zaradených škodlivých mutácii ku celkovému počtu škodlivých mutácii, teda koľko škodlivých mutácii bolo predikovaných správne.

[3] udáva pomer správne zaradených neutrálnych mutácii ku celkovému počtu neutrálnych mutácii, teda koľko neutrálnych mutácii bolo predikovaných správne.

[4] udáva precíznosť predikcie škodlivých mutácii, teda koľko predikovaných mutácii ako škodlivých je skutočne škodlivých.

[5] konsenzus z nástrojov AUTO-MUTE, MAPP a SIFT

[6] konsenzus z nástrojov AUTO-MUTE a SIFT

[7] konsenzus z nástrojov MAPP a SIFT

[8] konsenzus z nástrojov AUTO-MUTE a MAPP

V teste dopadol najhoršie nástroj AUTO-MUTE, ktorý mal najslabšie výsledky vo všetkých vypočítaných mierach (Tabuľka 3). V súlade s týmto pozorovaním, vyradenie AUTO-MUTU viedlo k významnému zvýšenie presnosti konsenzuálnej predikcie (Tabuľka 3, Cons2MS). Presnosť konsenzu nástrojov SIFT a MAPP je vyšší než presnosť jednotlivých nástrojov. Pre porovnanie sú v tabuľke Tabuľka 3 Presnosť predikcií doplnené aj výsledky konsenzu dvojice nástrojov AUTO-MUTE a SIFT (Cons2AS) a MAPP a SIFT (Cons2MS).

Výsledky testovania naznačujú potenciál meta-servera pre zvýšenie presnosti predikcií. K tomu je však potrebne vykonať optimalizácie súčasných nastavení meta-servera. Dôležitým krokom je nájsť tiež nastavenie AUTO-MUTU, taktiež i ďalších nástrojov, aby sa zvýšila presnosť jeho predikcie na čo najširšej škále proteínov. Možným riešením je pridanie váh k jednotlivým nástrojom tak, aby sa znížil vplyv menej presných nástrojov (ako bol v tomto prípade AUTO-MUTE) v konsenze. K nájdeniu optimálneho modelu pre výpočet konsenzu a vyhodnocovaniu jeho kvality poslúži Weka [44] – nástroj s algoritmami pre strojové učenie a dolovanie dát. K tomu je ale potrebné zhromaždiť experimentálne dáta pre väčší počet proteínov. V neposlednom rade k zvýšeniu presnosti meta-servera by mala významne prispieť plánovaná implementácia ďalších predikčných nástrojov.

6. Záver

Hlavnou myšlienkou tejto práce je oboznámiť sa s bioinformatickými pojmami, preskúmať možnosti predikcie vplyvu mutácie na funkciu proteínu, oboznámiť sa s predikčnými nástrojmi a s pomocou nadobudnutých znalostí navrhnúť unifikované rozhranie pre prácu s nástrojmi používanými pre analýzu a predpovedanie vplyvu mutácii v proteínoch. Výsledný návrh bol implementovaný v podobe meta-servera, ktorý zapuzdruje vybrané predikčné nástroje a servery, a získava od nich požadované dáta, ktoré ďalej spracúva. Aplikácia vytvára vlastnú predpoveď – konsenzus ostaných predpovedí, čím sme vytvorili nový predikčný nástroj.

Z výsledku tejto práce by mali ťažiť hlavne bioinformatici a proteínový inžinieri, ktorým meta-server uľahčí prácu so spomenutými nástrojmi – jednoduchým spôsobom im poskytne možnosť zadania výpočtov a získanie a porovnanie výsledkov jednotlivých nástrojov na jednom mieste. Služba LLWS, ktorá je súčasťou tejto práce, je už v praxi využívaná aj ďalším bioinformatickým nástrojom – HotSpot Wizardom. Veľký prínos práce bol taktiež čisto osobný. Zoznámil som sa so spôsobom práce vo výskumnom týme a získal cenné skúsenosti s tímovou spoluprácou – analýzou, návrhom a implementáciou v spolupráci so študentom inej vysokej školy. Rozšírila sa mi predstava o možnostiach využitia počítačovej techniky v praxi a v neposlednom rade som si prehĺbil znalosti Unixu.

Výsledky predbežných testov meta-servera dávajú poznať, že úsilie je smerované správnym smerom. Ďalším testovaním a nájdením optimálnej konfigurácie predikčných nástrojov a vyvážením konsenzu by malo viesť k žiadaným výsledkom.

Meta-server je aj naďalej vo vývoji, pracuje sa na pridaní ďalších predikčných nástrojov, úpravách užívateľského rozhrania, odstránení súčasných nedostatkov a na vylepšení a zdokonalení konsenzu. Návrh konsenzuálneho výsledku a jeho riešenie je jednoznačne zdĺhavý proces, počas ktorého bude nutné previesť veľa testov na experimentálnych dátach a ich vyhodnocovanie pomocou strojového učenia dolovania dát. Celkovo je budúca práca na projekte smerovaná k tomu, aby sa meta-server stal viac užívateľský prívetivým, presnejším a vierohodnejším predikčným nástrojom vplyvu jednonukleotidových mutácii na proteín a jeho funkciu. Aplikácia meta-servera je dostupná z adresy http://loschmidt.chemi.muni.cz/metasnp/.

Literatúra

[1] Wikipedia.org: elektronické stránky slobodnej a voľnej encyklopédie [Online] Dostupný na URL http://wikipedia.org/ (máj 2009).

[2] Genetika – Váš zdroj informací o genetike [Online] Dostupný na URL http://genetika.wz.cz/ (máj 2009).

[3] Alberts, B., et al.: Základy buněčné biologie, Espero publishing, 1998, s. 630, ISBN 80-902906-0-4.

[4] Virtuální svět genetiky 2 - principy molekulární genetiky [Online] Dostupný na URL http://old.mendelu.cz/~agro/af/genetika/vsg2 (máj 2009).

[5] Obrázok rozdelenia aminokyselín [Online] Dostupný na URL http://matcmadison.edu/biotech/resources/proteins/labManual/images/amino_000.gif (máj 2009)

[6] SNPs: Variations on a theme [Online] Dostupný na URL http://www.ncbi.nlm.nih.gov/About/primer/snps.html (máj 2009).

[7] Ng, P. C. a Henikoff, S.: Predicting the Effects of Amino Acid Substitutions on Protein Function, Annual Review of Genomics and Human, 7, 2006, s. 61-80.

[8] Shao, Z. a Arnold, F. H.: Engineering new functions and altering existing functions, Current Opinion in Structural Biology, 6, 1996, s. 513-518.

[9] Kast, P. a Hilvert, D.: 3D structural information as a guide to protein engineering using genetic selection, Current Opinion in Structural Biology, 7, 1997, s. 470-479.

[10] Kaur, J. a Sharma, R.: Directed evolution: an approach to engineer enzymes, Critical Reviews in Biotechnology, 26, 2006, s. 165-199.

[11] Chica, R. A., Doucet, N. a Pelletier, J. N.: Semi-rational approaches to engineering enzyme activity: Combining the benefits of directed evolution and rational design, Current Opinion in Biotechnology, 16, 2005, s. 378-384.

[12] Cvrčková, F.: Úvod do praktické bioinformatiky, Academia, 2006, s. 150, ISBN 80-200-1360-1.

[13] The EMBL Nucleotide Sequence Database [Online] Dostupný na URL http://www.ebi.ac.uk/embl/ (máj 2009).

[14] GenBank. NCBI [Online] Dostupný na URL http://www.ncbi.nlm.nih.gov/Genbank/ (máj 2009).

[15] DNA Data Bank of Japan [Online] Dostupný na URL http://www.ddbj.nig.ac.jp/ (máj 2009).

[16] Mizrachi, I. K.: Managing sequence data, Methods in Molecular Biology, 452, 2008, s. 3-27.

[17] Sayers, E. W., et al.: Database resources of the National Center for Biotechnology Information, Nucleic Acids Research, 37, 2009, s. D5-15.

[18] NCBI – Protein [Online]
Dostupný na URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein (máj 2009).

[19] Apweiler, R., et al.: UniProt: the Universal Protein knowledgebase, Nucleic Acids Research, 32, 2004, s. D115-D119.

[20] Uniprot [Online]
Dostupný na URL http://www.uniprot.org/ (máj 2009).

[21] NCBI - SNP [Online]
Dostupný na URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp (máj 2009).

[22] The Worldwide Protein Data Bank. the Worldwide Protein Data Bank [Online]
Dostupný na URL http://www.wwpdb.org/ (máj 2009).

[23] Berman, H. M., et al.: The Protein Data Bank, Nucleic Acids Research, 28, 2000, s. 235-242.

[24] Claverie, J-M., a Notredame, C.: Bioinformatics for dummies, Wiley Publishing, 2006, s. 436, ISBN: 978-0-470-08985-9.

[25] Basic Local Alignment Search Tool [Online]
Dostupný na URL www.ncbi.nlm.nih.gov/BLAST/ (máj 2009).

[26] Chenna, R., et al.: Multiple sequence alignment with the Clustal series of programs, Nucleic Acids Research, 31, 2003, s. 3497-3500.

[27] Edgar, R. C.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics, 5, 2004, s. 113-132.

[28] Care, M. A., Needham, C. J., Bulpitt, A. J., a Westhead, D. R.: Deleterious SNP prediction: be mindful of your training data!, Bionformatics, 23, 2007, s. 664-672.

[29] Sorting Intolerant From Tolerant [Online]
Dostupný na URL http://sift.jcvi.org/ (máj 2009).

[30] Ng, P. C., a Henikoff, S.: Accounting for human polymorphisms predicted to affect protein function, Genome Research, 12, 2002, s. 436-446.

[31] Stone, E. A. a Sidow, A.: Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity, Genome Research, 15, 2005, s. 978-986.

[32] Friedman, N., Ninio, M., Pe'er, I., a Pupko, T.: A structural EM algorithm for phylogenetic inference, Journal of Computational Biology, 9, 2002, s. 331-353.

[33] AUTO-MUTE [Online]
Dostupný na URL http://proteins.gmu.edu/automute/ (máj 2009).

[34] Masso, M. a Vaisman, I.I.: Accurate prediction of enzyme mutant activity based on a multibody statistical potential, Bioinformatics, 23, 2007, s. 3155-3161.

[35] Karchin, R.: Next generation tools for the annotation of human SNPs, Briefings in Bioinformatics, 10, 2009, s. 35-52.

[36] TORQUE PBS [Online]
Dostupný na URL http://www.clusterresources.com/products/torque-resource-manager.php. (máj 2009)

[37] Pavelka, A., Chovancová, E. a Damborsky, J.: HotSpot Wizard: a web server for identification of hot spots in protein engineering, Nucleic Acid Research: Web Server Issue, 37, 2009.

[38] Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., a Ben-Tal, N.: Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, Bioinformatics, 18, 2002, s. S71-S77.

[39] Expo|CMS [Online]
Dostupný na URL http://expocms.net/ (máj 2009).

[40] Rennell, D., et al.: Systematic mutation of bacteriophage T4 lysozyme, Journal of Molecular Biology, 222, 1991, s. 67-87.

[41] Loeb D.D., et al.: Complete mutagenesis of the HIV-1 protease, Nature, 340, 1989, s. 397-400.

[42] Wrobel, J.A., et al.: A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase, Proceedings of the National Academy of Sciences of the United States of America, 95, 1998, s. 638-645.

[43] Kawabata, T., Ota, M. a Nishikawa, K.: The Protein Mutant Database, Nucleic Acids Research, 27, 1999, s. 355- 357.

[44] Weka 3 - Data Mining Software in Java [Online]
Dostupný na URL http://www.cs.waikato.ac.nz/ml/weka/ (máj 2009)

Zoznam príloh

Príloha č. 1. Konfiguračné súbory LLWS pre nástroje AUTO-MUTE, MAPP a SIFT.
Príloha č. 2. CD so zdrojovými kódmi meta-servera, návodom k inštalácii a generovanou programovou dokumentáciou.

Príloha č. 1

Konfiguračný súbor pre Automute v LLWS

```
[command]
command = "echo '#PBS -N MS_AUTOMUTE_${output}
php ${path}/automute.php ${output} $pdbId $chain $subst' | cat | qsub"
[obligatory]
Data = "pdbId;chain;subst"
[check]
check = "${path}/results/${output}.result"
[const]
path = "/usr/aas/sw/automute"
```

Konfiguračný súbor pre MAPP v LLWS

```
[command]
command = "echo '#PBS -N MS_MAPP_${output}
php ${path}/mappHandler/mappHandler.php ${output} $fasta $align' | cat | qsub"
[obligatory]
File = "fasta"
Data = "jobMethodId"
[optional]
File = "align"
[check]
check = "${path}/mappHandler/results/${output}.xls"
[const]
path = "/usr/aas/sw/mapp"
[results]
msa="${path}/mappHandler/MSA/${output}/msa.fas"
```

Konfiguračný súbor pre SIFT v LLWS

```
[command]
command="echo '#PBS -N MS_SIFT_${output}
${path}/sift/sift3.0/bin/SIFT_for_submitting_fasta_seq.csh $fasta
${path}/blast/blast-2.2.19/data/nr90.fsa $subst' | cat | qsub"
[obligatory]
File = "fasta;subst"
[check]
check="${path}/sift/sift3.0/tmp/${output}.SIFTprediction"
[const]
path = "/usr/aas/sw/tools"
```